Preface

MÁRIA RAFFAI
Editor in Chief; Chair of GIKOF/SEFBIS SIG
Professor at Széchenyi University, NR in and Councilor of IFIP,
vice chair of IFIP Information Science TC Enterprise Information Systems WG
eMail: maria.raffai@gmail.hu

After that the Scientific and Educational Forum for Business Information Systems (SEFBIS) as a Special Interest Group of the John von Neumann Computer Society was established in 2000 the founders decided to launch a conference series on BIS. Later on, at one of the first conferences the participants expressed their wish for publishing the abstracts of their papers not only in a booklet, but also in a citied professional journal. As our conferences turned to be international, there emerged a need for publishing journals in English. The members of the SEFBIS/GIKOF Board made all the efforts to find supporters, reviewers, solutions for editing and printing the Journals and last but not least to get sponsors for financing the edition. If we look back on the last years' activity we can be proud of our results, namely the 12 Volumes of SEFBIS Journal (in English) and the 11 Volumes of GIKOF Journal (in Hungarian). These publications contain 8-10 papers each in wide variety of topics from scientific results through business application to professional trainings. Both the IT experts and the users appreciated our efforts and the enormous volunteer work we do for giving opportunity for scientists, professionals, developers and users to publish their results.

The present volume of SEFBIS Journal contains 8 papers mainly from three different fields of IT:

1. *Scientific results*
- Implementing and evaluating different machine learning algorithms
- Random correlation methodology as a tool for analyzing random factor of scientific results
- Forecasting performance improvement: comparison automated statistical and neural network-based methods

2. *Data management solutions*
- A beginner's guide to open data – A case study
- Big data analytics possibilities for the space domain

3. *ICT in business*
- Measuring organizational information security awareness – Levels supported by a maturity model
- The connection between the production and the energy usage in a smart factory
- Business information visualization in tangible ways

This Journal is also reported on the BIS Community's main conferences: the OGIK and CONFENIS organized in the last year and called the attention to the 2018 and 2019 events the readers might be interested in.

Performing our goals also in the future we call the teachers, the professionals, the developers to report their results, the efficient business and/or educational solutions. As we know from the conferences, from the forums and discussions the readers are interested in

- the role and the impact of IS/IT on business and on society,
- the concepts, modeling techniques, methods, visualization languages supporting the IS development processes,
- the solutions that satisfies customers' requirements and ensures security and privacy, and
- the realization of the Higher Education Space focusing to the field of business information systems (aims, programs, curricula, cooperation, new teaching materials etc.)

Concluding with my sincere greeting to the Readers I whish to obtain new ideas, knowledge about effective ICT innovations in business and research results from all over the world! I call the attention of the specialists in the field of business information science to make their results public and well known internationally!

The electronic version of the Journal is citied and downloadable from EBSCO Database and from the SEFBIS HomePage: http://raffa6.wix.com/sefbis#!journals

Implementing and Evaluating Different Machine Learning Algorithms to Predict User Localization by the Strength of User Devices' Wi-Fi Signal

GERGŐ BARTA
PhD student, Szent István University
eMail: Barta.Gergo@phd.uni-szie

ABSTRACT

The objectives of the article is to implement 6 different machine learning algorithms and evaluate their performance which aim to predict appropriate user localization in indoor environment based on the signal strength that is produced by their smartphones. The field of application area is diverse. Such an application established on the concept of machine learning can be used to target customers real-time in shopping malls with direct marketing and sales, trace users in office areas for further research of their working habit or count the users in specific rooms for statistical purposes. The article is going lead us through the description, implementation and evaluation of selected algorithms which are logistic regression, support vector machines, K-nearest neighbors, decision tree, neural network and K-means clustering. As hybrid solutions, two different combined models have been created from the utilized algorithms to take advantage of the different characteristics of the methods and increase the prediction power.

Introduction

The localizations by Wi-Fi signal strength provides good opportunities. Similar projects can be imagined by installing microchips under the skin of users, however, its culture is not widely-spread yet and such solution needs further investment in radio frequency receivers, GPS devices can provide poor quality inside of buildings because of the weaker satellite signals, and bluetooth technology can only serve in a determined range [1]. Another problem with indoor localization techniques based on technologies rely on signals that they might be reflected of walls and metals. Nevertheless, beside the Wi-Fi signals, Ultra-wideband or Visible Light could also be utilized to give a prediction of localization in indoor environment, in addition, if the hot-spot coordinates and/or building plans are available, definitely better results can be achieved of spotting users over only using Wi-Fi signal strength.

In order to give a prediction of users' whereabouts in indoor environments, machine learning can provide a wide range of set of algorithms. For such a purpose, historical data collection is needed. Based on the available data, the application is able to give an estimation in case a user is similar to another and predict the room the user stays. We can interpret this task as a classification problem, where the features are the strength of signals from different routers and the target variable to predict is the coordinate or the identifier of specific room. Supervised learning, a type of machine learning, can be used to predict a class based on such data previously collected by identifying patterns in known data or training data and draws conclusions on unseen data or testing data. So as to find the best-to-fit classifier, an error function should be defined which is to be minimized, but care should be taken to avoid overfitting that occurs if the classifier fits very well on the training data, but performs low on test data. Therefore, in the algorithms used for this experiment, the whole data set has been randomly divided by training and test data in a 70% and 30% ratio, and the evaluation has been performed only on the test data to inspect the prediction power of the algorithms. In addition, an unsupervised method, K-means clustering, was also utilized as

the sixth model to understand the nature and extent of the data set and perform prediction for the target variable by observing whether the algorithm could group appropriately the records as they originally appeared in the data set. Since K-means is an unsupervised learning method, therefore, the target variable was not relevant to be involved in the analyzing process for this model. The algorithms used purposely are the simplest methods with no hybrid solution involved established on the idea of Occam's razor, as among the solutions, always the simplest is preferred [3].

The data set consists of 2.000 records, i.e. the descriptive data of 2.000 users with 7 individual and independent features which represent 7 different routers placed in 4 different rooms, therefore the target variable can take 4 different values. The data set was collected in an office location in Pittsburg, USA and is provided by Lichman (2013) and can be downloaded from the UCI Machine Learning Repository [4]. Few lines of samples from the original data set are shown on Figure 1. The features' measure scale is dBm (decibel-milliwatts) which is used to measure the Wi-Fi signal strength of a device connected to a router. Not to dig deeper into the mathematical background, enough to understand that Wi-Fi signal strength is measured in a negative scale, however, the values originally represent very small positive values measured on a logarithmic scale, e.g. under the natural logarithm 0.1 results in the value of around -2.3. Thus, a negative value of dBm represents that a negative exponent is applied. Based on the equation used in calculating dBm [5], -10 dBm equals 0.1 mW, -20 dBm equals 0.01 mW and so on. The larger the dBm the better the quality of the Wi-Fi, the signal strength. With this information the data set can be interpreted. In case of the first sample, the signal is the best for Router 2, then Router 3, then Router 1 etc. To summarize, the main task is to predict the Room Identifier based on the measured Wi-Fi signal strength with the algorithms intended to be used for the experimentation.

Router 1	Router 2	Router 3	Router 4	Router 5	Router 6	Router 7	Room Identifier
-64	-56	-61	-66	-71	-82	-81	1
-68	-57	-61	-65	-71	-85	-85	1
-37	-51	-54	-43	-60	-69	-70	2
-36	-56	-57	-46	-68	-68	-70	2
-51	-52	-56	-56	-63	-79	-80	3
-51	-54	-52	-57	-62	-79	-79	3
-64	-57	-53	-64	-47	-88	-91	4
-64	-59	-51	-62	-51	-88	-91	4

Figure 1. Sample data from the original data set [4]

Related work

Bozkurt et al. [6] developed several machine learning algorithms to predict indoor positioning by analyzing Received Signal Strength (RSS) values from 520 wireless access points and related building information. In the experiment, the problem was similarly defined as a classification problem divided into three phases. In the first phase, the authors predicted the building label, in the second phase, the floor label, and in the last phase, the region label was determined. The used algorithms were K-nearest neighbours (KNN), Sequential minimal optimization (SMO), decision tree (J48), Naive Bayes and Bayes Net. In case of the decision tree, AdaBoostM1 and Bagging were utilized to improve the performance. The performance of each classifier was evaluated by comparing the accuracy and computational time of the algorithms. Regarding the first phase, each of the algorithms performed above 99% of accuracy in the order of Bayes Net (99.8%), SMO and KNN (99.7%), decision tree (99.4%) and Naive Bayes (99.2%). In the second phase, the KNN algorithm performed the best accuracy with an

average of 98.5% followed by the decision tree (98%), SMO (95.1%), Bayes Net (95%) and the Naive Bayes classifier (67.4%). In the last phase the KNN algorithm had again the best prediction power and SMO was the second one. As a comparison of each of the algorithms, all over the KNN provided the highest accuracy and it required less execution time. Rohra et al. [1] in their study, examined the same data set used in this article, however, only three different indoor locations were in their scope. They utilized four different neural network models (Particle Swarm Optimization-NN i.e. PSO-NN, Gravitational Search Algorithm-NN i.e. GSA-NN, a hybrid model of the first two algorithms that is PSOGSA-NN and FPSOGSA-NN that is a hybrid model of the first three algorithms involving fuzzy logic) and compared their results to support vector machines and Naive Bayes. They measured the classification rates of the algorithms and concluded that the best performance was achieved by the FPSOGSA-NN with 95.2% accuracy followed by the support vector machine (92.7%), Naive Bayes (90.5%), PSOGSA-NN (83.3%), GSA-NN (77.5%) and the PSO-NN (64.7%). Zafari et al. analyzed the current technological background of indoor localization systems and proposed a framework based on different metrics (such as energy efficiency, accuracy, scalability, reception range, cost, latency and availability) that can help evaluate indoor localization techniques. The authors highlighted that with the rise of Internet of Things (IoT) indoor localization became in the focus of interest among researchers, since it can provide a wide range of services. They provided detailed description of different techniques such as Angle of Arrival, Time of Flight, Return Time of Flight, Received Signal Strength Indicator and technologies such as Wi-Fi, Ultra-wideband, Visible Light etc.

Description of the algorithms

Logistic regression

Logistic regression is a statistical method which aims to predict the probability of a given target variable based on the inputs known as feature vectors. Logistic regression by itself can be useful to predict a binary variable. Let us call the target variable y, therefore y€{0,1} where 0 represents the negative event and 1 represents the positive event e.g. predicting whether there will be rain tomorrow or not. Let's introduce p as the probability of a certain event where 0≤p≤1. We can interpret it as if p>0.5 then we classify our target variable as 1, otherwise 0. In order to estimate the value of p we shall introduce a function $f(X)=1/(1+e^x)$ which outcome is in our range where X is the net input in the function. This function can be the sigmoid function which received its name based on its distinctive S-shape as shown on Figure 2.

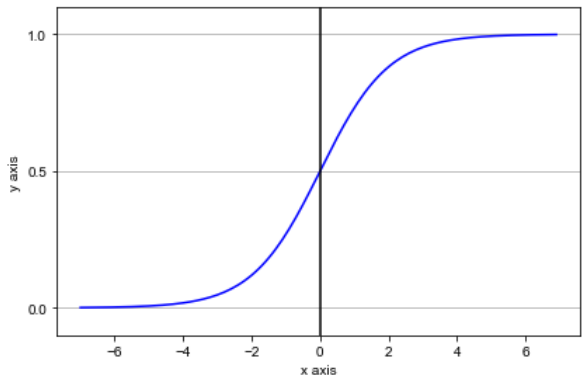


Figure 2. The sigmoid function

In a classification problem, where n is the number of features, x are the features, and w are the weights, X can be calculated as $X=w_0x_0+w_1x_1+\ldots w_nx_n$. Thus, the output of the function is the probability of a feature vector belonging to a given class, where the class label is c, $p=f(X)=P(y=c|\ x;w)$. In our case, we have 4 different classes. The good news is that there are techniques such as One-vs-All (or One-vs-Rest) that can be used in multiclass classification problems. It means that we have to train one single classifier in case of each classes e.g. for y=1, we deem only the first class as positive, and the rest as negatives. We repeat it for each class, calculate the corresponding probabilities for the classifiers, and choose the one with maximum value out of the results e.g. in case of the following [0.35, 0.55, 0.05, 0.05] [0.35, 0.55, 0.05, 0.05] we conclude that the predicted class is the second one.

Support vector machines

The generic idea behind support vector machines is to maximize the margin between the decision boundary which separates the training data belonging to different classes and the samples. One of the biggest reasons to use this algorithm is to avoid overfitting, since large margins are less exposed to the phenomenon, the data samples from different classes are the farthest from the decision boundary.

Figure 3. illustrates the outcome decision boundary by utilizing support vectors. On the left side of the figure, we can observe that many decision boundaries can exist which separates the data differently e.g. by using logistic regression, however, on the right side, only one decision boundary can be seen where the margin is maximized and placed in the middle of the two classes.

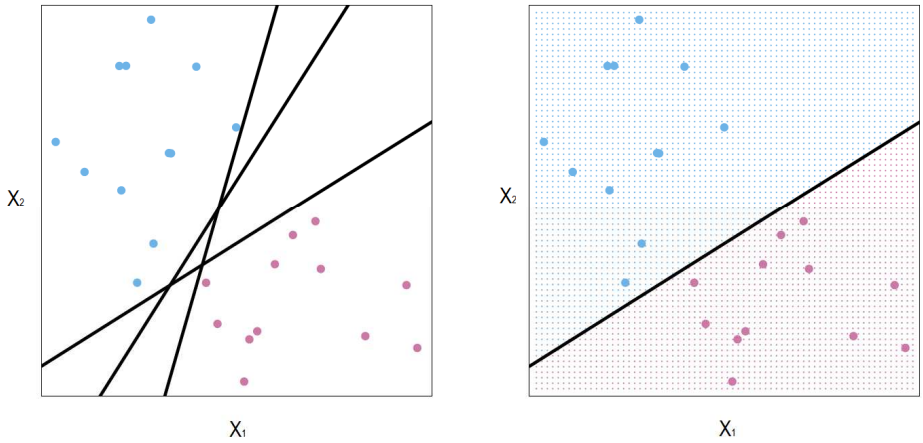


Figure 3. Maximizing margin with support vectors [7]

The samples placed the nearest to the decision boundary are called the support vectors. When the data set cannot be separated linearly, support vector machines can be kernelized. Kernels are to help on the problem by finding separators in a higher dimensional space which behaves as a similarity function between the samples. Since support vector machines are designed to maximize the margin between the classes and thus, avoid overfitting, the presumption is that the algorithm performs well among the models considering the characteristics of the data set.

K-nearest neighbors

K-nearest neighbors algorithms work by majority voting which means that a value of k has to be determined, and in case there is a new sample a class label is chosen based on the k closest samples from a given class e.g. if k=3, and the 3 closest samples are classified as [1,1,0] we predict the class label to be 1 for the unseen sample since the majority of the samples belonged to class 1. The best is to set an odd number for the parameter k, since the equal number of class labels e.g. 50-50 ratio can be avoided. In addition, a distance metric shall be determined to appropriately make a decision which are the k samples being the closest to the unseen data. The algorithm can be even more fine-tuned if we consider that samples are rarely close to each other in the same proximity, therefore proportional calculation can be applied and we can assign the greater weights to the samples which are the closest to the data to be predicted. Because of the spatial structure of the data used in here (proximity in a geographical space), the KNN is expected to be one of the algorithms with high accuracy.

Decision trees

In decision tree models we can train the algorithms to establish a series of questions from the training data and classify the test data based on the trained tree. Figure 4. serves a simplified example to decide a Sunday evening activity.

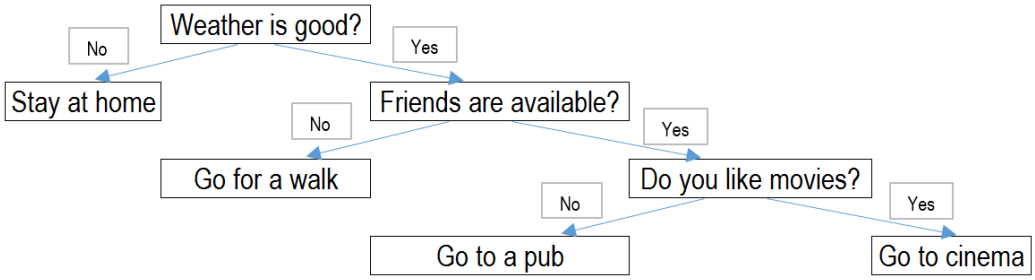


Figure 4. Decision tree example

In order to decide how to split the tree starting from its root a method shall be determined which can be e.g. the Information Gain calculating which features and its corresponding values can best split the tree into sub-decisions iterating the process until the target variable is pure i.e. we can be confident which class label to choose for an unseen sample. So as to avoid the overfitting problem we can set e.g. the minimum number of samples a leaf must contain, the maximum depth or the ratio in the distribution of class labels in the leaves. Decision trees generally generate high-variance models, therefore they are prone to overfitting. The presumption is that in case of the models, the decision tree will perform one of the lowest.

Neural networks

Neural networks are algorithms which were inspired by human brain cells first published by Warren McCullock and Walter Pitts [8]. Ever since neural networks became so popular that as of now deemed to be a leading solution for researching the possibilities of use of autonomous vehicles, speech and image recognition and so on. The basic idea behind neural networks is to model the operation of how the human brain works. Neural networks are based on several connected neurons which transfer a signal from one neuron to another one. The receiver neuron processes the signals and then sends them towards. Processing means that a neuron collects each of the signals multiplicated by the weights, which actually contains the „knowledge" and by an activation function, calculates the output. Neural networks can be used for supervised and unsupervised learning as well. A neural network consists of an input layer, initiating the feature vectors, hidden layers, processing the signals, and an output layer, providing the predictions. A neural network can be made deeper by increasing the hidden layers, and it can contain many nodes, therefore, countless number of architecture can be created. One of the most popular algorithm to update its weight is gradient descent which is an iterative method for finding the minimum of the objective function of the network. The convergence of a neural network can be regulated by the learning rate to avoid the overfitting of the network. Figure 5. illustrates a neural network containing 5 nodes of the input layer, 3 hidden layers with 6, 7, and 6 nodes, and an output layer with 5 nodes.

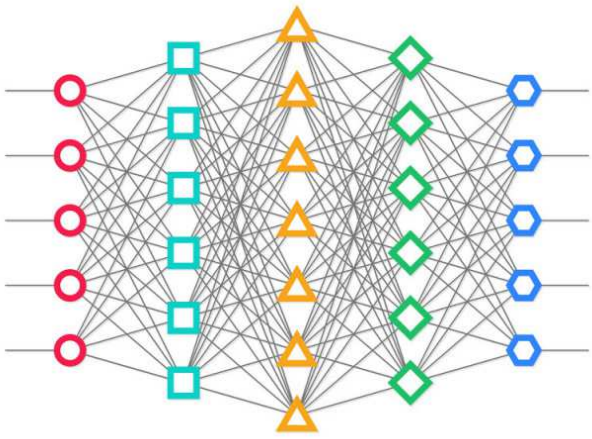


Figure 5. A neural network [9]

K-means clustering

Clustering is a method that helps to find groups of records in a data set based on their features' similarity. K-means is an algorithm where the number of clusters should be defined in advance, therefore, if the choice for the number of clusters is not ade-

quate, then the method can result in poor performance. In this experiment, this problem cannot cause any headaches, since it is already known that 4 clusters are needed, since there are 4 rooms from where the data has been collected. K indicates, in the name of the algorithm, the number of clusters chosen a priori, thus, in this case it can be called 4-means clustering. The number of iterations can be determined for the algorithm that indicates after how many repetitions it should stop.

Since the data set can be derived into a two dimensional subspace, considering that it is a problem to be solved in a geographical space, the basic expectations is that the algorithms will perform quite well all over on the test set.

Majority voting

In order to make the most out of the models, two hybrid solutions were programmed to achieve a better prediction results. The first one is called Majority Voting which means that we predict the class label that has been selected by the majority of the algorithms. The only problem is that it is possible that there is no winner, and the game is tie between two methods if we have different predictions twice e.g. in case the algorithms predict the following values [logistic regression = 1, support vector machines = 2, K-nearest neighbour = 2, decision tree=1, neural network=3, K-means cluster = 4] i.e. two algorithms voted 2, and two algorithms voted 1, then we have to decide which is the winner. The following rule has been established: wait until we predicted each value and chose the one which has been voted more times than the other. So in case we have a prediction of class label 2, 130 times, and we have a prediction of class label 1, 150 times, then the algorithm will choose the class label 1.

Hybridization by probabilities

The second hybrid model is formulated on the basis of probabilities that the supervised learning methods produce in each case of the testing set samples. Since the K-means algorithm does not give an estimation of the probability of belonging one sample to a class, on the contrary of the rest of the algorithms, therefore it was not involved in the second hybrid model. For example, if the logistic regression predicts in case of the first sample that it has the probability belonging to the first class with 97.1%, to the second class with 0.001%, to the third class with 2.91%, and to the forth class with less than 0.001% chance, the predicted class label becomes 1, as it has the highest probability. These values are weighted by the accuracy of the individual models and then they are summarized for the five models and divided by the sum of the models' weights so as to obtain a value between 0 and 1. After the computation for each class, the maximum value is selected that indicates the predicted class label of given sample, i.e. for one feature vector:

$$\hat{y} = MAX \left(\frac{\sum^{5}_{m=1} W_m * P(y=1)}{\sum^{5}_{m=1} W_m} , \dots \frac{\sum^{5}_{m=1} W_m * P(y=4)}{\sum^{5}_{m=1} W_m} \right)$$

where $\hat{y}$ is the predicted label, W_m is the weight of particular model and P(y=1) is the probability of class label 1. Since this model is more sophisticated than majority voting, as it is taking into consideration the probabilities of samples belonging to a given class, the presumption is that this model is going to perform the best on the testing set.

Implementation

The following steps were performed in case of the selected algorithms:

1. Read the file containing the data set into the interpreter.
2. Separate the feature vectors, load them to a matrix and the class labels to a vector.
3. Divide the data set into training and test set in a 70%-30% ratio.

4. Standardize the training set and test set to achieve better performance.
5. Build the model and set corresponding parameters.
   - Support vector machines: kernel = linear.
   - K-nearest neighbors: k = 3, distance metric = Minkowski.
   - Decision tree: Criterion=entropy, max depth=3.
   - Neural network: activation function = RELU, learning rate = 0.0001, hidden layers = 2, nodes per hidden layers = 5.
   - K-mean clustering: k=4, maximum iterations=1000
6. Evaluate the model performance on the test set.
7. Evaluate the model performance inspecting the learning curves.
8. Evaluate the model performance analyzing the confusion matrices.

Evaluation of selected algorithms

The summary of the performance of utilized models is shown on Figure 6 and Figure 7. The results can only be interpreted on the separated 30% test data, which are all over 600 samples.

As we can observe, there is not much difference between the prediction power of different supervised algorithms, however, the best performing was the K-nearest neighbors with 97.67% accuracy, which means that it perfectly predicted 586 room identifiers out of the total 600. The deviation is negligible among these methods, since the worst performance was produced by the decision tree, but it could still appropriately predict 578 samples, thus the difference between the winner and the worst performance is 8 samples, i.e. 1.33%. The K-means clustering has performed the lowest with 92.83% accuracy that is 557 correct labels. Comparing the results of the work of Rohra et al. [1] on the same data set, except the K-means clustering algorithm, each of the utilized supervised methods performed better accuracy than the best model (FPSOGSA-NN with 95.2%) in the referenced article.

Figure 8. represents the learning curves of the algorithms. The learning curves show that how the prediction power has changed as the algorithms had consumed more and more training samples and modified their weights as they learnt more from the pattern. The learning curves cannot be interpreted for the K-means algorithm as in its case, there was no learning phase.

The upper line indicates the prediction power of the training set, the green indicates the prediction power of the test set. The logistic regression, support vector machines and K-nearest neighbors algorithms have already achieved the maximum prediction power after 200 processed samples and there are no noted changes after, thus, around 200 samples were enough to be processed by the algorithm to achieve the intended goal and give a good prediction about the users' whereabouts. The decision tree is very similar, slight change can be inspected after 1.000th processed sample. The neural network behaves a bit differently. Its maximum prediction power was achieved after 800 samples, and its performance was zero after 300 processed samples due to its slowlier convergence.

Logistic regression	Support vector machines	K-nearest neighbors	Decision tree	Neural network	K-means clustering
97.17%	97.50%	97.67%	96.33%	96.67%	92.83%

Figure 6. Evaluation of prediction performance on the test set I.

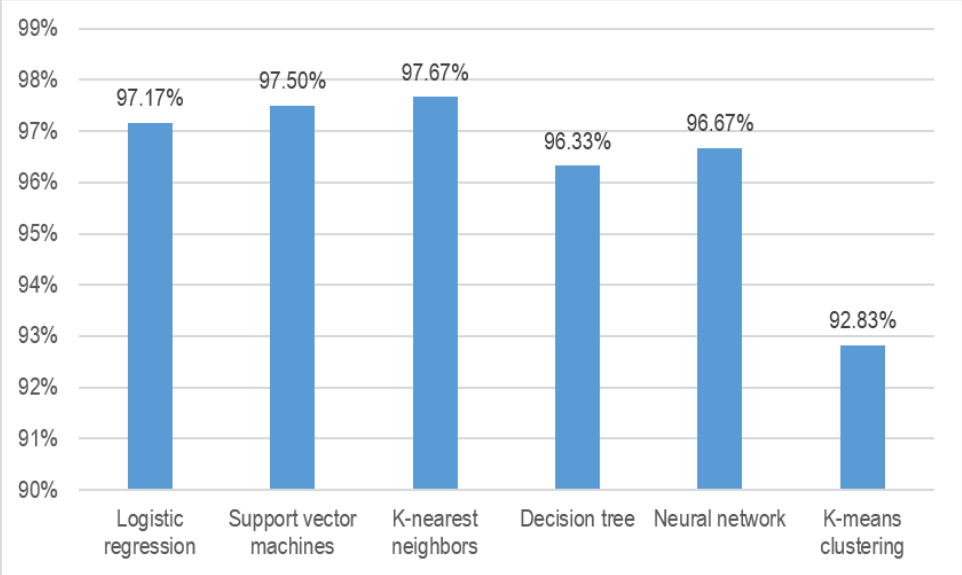


Figure 7. Evaluation of prediction performance on the test set II.

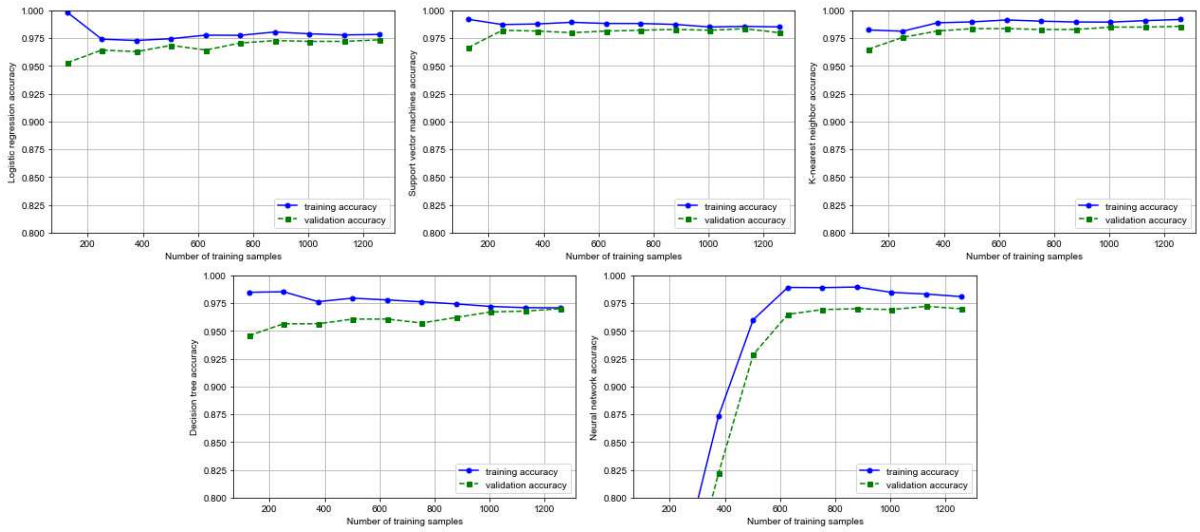


Figure 8. Learning curves of the algorithms

Evaluation of the hybrid models

The first developed hybrid model was the majority voting. After executing the voter, a 98.17% accuracy has been achieved on the testing set which is only a slightly better than the best performing algorithm, with 3 samples of better classification. In case of 2 samples, the voter decided to select the most popular class label, and predicted it wrongly in 1 case.

Figure 9. represents the confusion matrices of the algorithms including majority voting. The matrices report on the exact prediction of different samples and in case any false result, it shows which label was predicted instead. E.g. the logistic regression was able to predict correctly 156 samples which belonged to the Room 1, but in one case, it predicted wrongly that a sample belonged to Room 3, instead of Room 1. The rest of the lines can be interpreted similarly.

True Classes:

	Logistic regression				Support vector machines				K-nearest neighbor			
	1	2	3	4	1	2	3	4	1	2	3	4
1	156	0	1	0	157	0	0	0	157	0	0	0
2	0	147	4	0	0	140	11	0	0	141	10	0
3	0	8	142	3	1	1	150	1	1	2	150	0
4	0	0	1	138	1	0	0	138	0	0	1	138

	Decision tree				Neural network				K-means clustering			
	1	2	3	4	1	2	3	4	1	2	3	4
1	155	0	2	0	156	0	1	0	142	0	15	0
2	0	140	11	0	7	137	7	0	3	131	17	0
3	0	3	149	1	2	0	150	1	4	1	146	2
4	2	0	3	134	0	0	2	137	0	0	1	138

	Majority voting			
	1	2	3	4
1	156	0	1	0
2	0	143	8	0
3	0	0	152	1
4	0	0	1	138

Predicted Classes

Figure 9. Confusion matrices of the algorithms

The hardest job was to predict Room 2 for the algorithms, except the logistic regression which has performed poorly on Room 3. Room 1 and Room 4 prediction occurred definitive easily, the biggest problem was caused by Room 2 and Room 3 which were mixed up a few times by the algorithms.

True classes	Hybridization by probabilities			
	1	2	3	4
1	157	0	0	0
2	0	144	7	0
3	1	0	151	1
4	1	0	0	138

Predicted classes

Figure 10. Confusion matrix of the second hybrid model

After the execution of the second hybrid model, the prediction power has increased by 0.16% (as of now 98.33%), i.e. 1 sample plus, however, all over 4 sample difference can be inspected on the confusion matrix shown on Figure 10. We can see for 7 samples, Room 3 was similarly predicted in most cases when mistakes occurred, instead of Room 2.

It is interesting to analyze the 7 samples that were mistaken by the majority of the algorithms, as well as the hybrid models. Figure 11. shows the Euclidean distance of the 7 samples calculated against the average of Class 2 and Class 3, and we can observe, that each of the cases the 7 samples were closer to Class 3 than Class 2 indicating, that most probably, even though the users stayed in Room 2, they were closer to the routers of Room 3, on an average basis. Since the accuracy is 98.33%, we can say with a high confidence, that the algorithm could appropriately separate which users stayed in which rooms.

	Class 2 (AVG)	Class 3 (AVG)
1	18.63	8.51
2	16.74	16.21
3	17.44	12.53
4	14.00	13.12
5	18.25	9.91
6	18.67	10.15
7	18.29	15.45
Class 2 (AVG)	0	21.26
Class 3 (AVG)	21.26	0

Figure 11. Euclidean distance of the 7 misclassified items and the average of class 2 and class 3

Conclusion

This experiment has proven that localizing users in indoor environment by the strength of user device Wi-Fi signals can lead to promising results. As we could observe, each of the utilized machine learning algorithms were capable to predict the target variable i.e. the room of specific users located around 96%-97% well which opens many opportunities for organizations to take advantage of the collected data and analyze indoor activities for business purposes. As the continuation of the article, it is worth considering to analyze the feature vectors deeper which caused the headaches for the algorithms and build up new hybrid models to achieve an even better prediction performance.

Acknowledgement

To express my gratitude towards the contributor of the data set, Rajen Bhatt, I would like to recommend his book [10] titled as *Fuzzy-Rough Approaches for Pattern Classification: Hybrid measures, Mathematical analysis, Feature selection algorithms, Decision tree algorithms, Neural learning, and Applications*.

References

[1] Jayant, G, Rohra, Boominathan, Perumal, Swathi, Jamjala Narayanan, Priya, Thakur, and Rajen B Bhatt (2017). User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks. In Proceedings of Sixth International Conference on Soft Computing for Problem Solving, pp. 286-295.

[2] Zafari, Faheem, Gkelias, Athanasios, Leung, Kin, K. (2018). A Survey of Indoor Localization Systems and Technologies. Available: https://arxiv.org/pdf/1709.01015.pdf. [Accessed: 28.05.2018]

[3] Pitlik, László (2014). Occam hermeneutikája. Magyar Internetes Agrárinformatikai Újság. Available: http://miau.gau.hu/miau2009/index.php3?x=e0&string=occam. [Accessed: 14.01.2018]

[4] Lichman, M. (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science. [Accessed: 16.01.2018]

[5] ISA publications (2002). dB vs. dBm. Accessed: https://www.isa.org/standards-publications/isa-publications/intech-magazine/2002/november/db-vs-dbm/. [Accessed: 15.01.2018]

[6] Bozkurt, Sinem, Elibol, Gulin, Gunal, Serkan, Yayan, Ugur (2015). A comparative study on machine learning algorithms for indoor positioning. In Proceedings of International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), pp. 47-55.

[7] Tendel, Aakash (2017): Support Vector Machines – A Brief Overview. Available: https://towardsdatascience.com/support-vector-machines-a-brief-overview-37e018ae310f. [Accessed: 19.01.2018]

[8] McCullock, Warren and Pitts, Walter (1943): A Logical Calculus of the Ideas Immanent in Nervous Activity. The bulletin of mathematical biophysics, 5(4). pp. 115-133.

[9] Pangeanic.com (2018): What are machine translation neural networks? Available: https://www.pangeanic.com/translation-technology/machine-translation/what-are-machine-translation-neural-networks/. [Accessed: 22.02.2018]

[10] Bhatt, Rajen (2017). Fuzzy-Rough Approaches for Pattern Classification: Hybrid measures, Mathematical analysis, Feature selection algorithms, Decision tree algorithms, Neural learning, and Applications. Amazon Books.

Random Correlation Methodology as a Tool for Analyzing Random Factor of Scientific Results

GERGELY BENCSIK
Sopron University
eMail: bencsikg@inf.uni-sopron.hu

ABSTRACT

Nowadays, there are many data measured by sensors and analyzed by various models and methods. The process is also supported by Big Data and Internet of Things phenomena. These environments extend the possibilities to find correlations that are more precise. However, based on our experiments, these constantly increasing possibilities can create conditions, near which the results are born just randomly, despite of the followed exact mathematical environment. The random property of the result is hidden from the scientists. In this paper, a new process is introduced, with which the random factor level of the given result can be determined.

Introduction

There are lots of possibilities to collect data nowadays. Therefore, standard research methodologies are defined by many state-of-art publications [1, 2]. Specialized research methodologies also appear corresponding to the given research fields [3, 4]. In general, an analyzing session starts with the data preparations (collect, clean and/or transformation), continues with choosing an analytical method and finally, the result is presented and interpreted. If we have a lot of data items, we talk about Big Data, which can provide more analyzing possibilities and more precise results, as we would expect. But a lot of contradictory results were born in different scientific fields and the literature contains many inconsistent statements.

In biology, Zavaleta et al. stated that grassland soil has more moisture [5]. According to Liu et al., grassland soil must face against less moisture [6]. Church and White showed out a significant acceleration of sea level [7]. However, comparing the results with [7], Houston and Dean results show us see-level deceleration [8]. In medicine, analyses of the salt consumption are always generating opposite publications. There are papers supporting the consumption and do not disclose any connection between consumption and high blood pressure [9]. Another research group states that the high salt consumption causes not only high blood pressure but kidney failure as well [10]. In forestry, Held et al. have stated that Sahel, a transition zone between Sahara and savanna in the north part of Africa, has less rain [11]. However, another research group suggested more rain for Sahel [12]. In Sahel local point of view, Giannini's result was that it may get more or less rain [13]. In sociology, analyses of Massively Multiplayer Online Role-Playing Games (MMORPG) also generate contradictory results. Brian and Wiemer-Hastings stated that MMORPG can cause addiction [14]. However, Yee result stated that video games have different consequences for different players and they do not always cause addiction [15]. In Earth science, Schindell et al. stated that winters could getting warmer in the northern hemisphere [16]. According to other opinion, winters are maybe going to colder there [17]. Knippertz et al. deal with wind speeds and they concluded that wind speed become faster [18]. Another resource group stated that wind speed is declined by 10-15% [19]. According to the third opinion, the wind speed speeds up, then slows down [20]. Nosek et al. repeat 98 + 2 psychology researches (two were repeated by two individual group) [21]. Only 39% of the publications showed the same significant results as before. In other words, contradictory results, between the repeated researches and the original researches, came out.

The above-mentioned researches focus on the same topics, but they have different results. This shows us how difficult the decision making could be. Our research focuses on how the inconsistent results could be originated. This does not mean that one given problem cannot be approached with different viewpoints. We state that there are circumstances, when the results could have born due to simple random facts. In other words, based on parameters related to data items (e.g., measured items range, mean and deviation) and methods of analyses (e.g., number of methods, outlier analysis) can create such environment, where the possible judgment is highly determined (e.g., data rows are correlated or non-correlated, pendent or independent). Based on our results, a new phenomenon named Random Correlation (RC) is introduced.

Random Correlation Framework

In this section, Random Correlation Framework (RCF) and its components are presented. First, the precise definition of RC is introduced. Then, parameters, which is used to describe data structure, will be presented. Finally, we deal with the calculation models, with which RC can be measured.

Definition

The main idea behind the Random Correlation is that data rows as variables present the revealed, methodologically correct results, however these variables are not truly connected, and this property is hidden from researchers as well. In other words, the random correlation theory states that there can be connection between data rows randomly which could be misidentified as a real connection.

There are lot of techniques to measure result's endurance, such as R^2, Bootstrap, Cramer Coefficient or Type I and II errors etc. We do not intent to replace these measurements with RC. The main difference between "endurance measurement" values and RC is the approach of the false result. If we have a good endurance of the result, we strongly assume that the result is fair, or the sought correlation exists. RC means that under the given circumstances, we can get results with good endurance. We can calculate R^2, critical values and Type I and II errors, we can make the decision based on them, but the result still can be affected by RC. From the RC point of view, the main question is that if we have the set of the possible inputs, how the result can be calculated at all.

Parameters

Every measured data has its own structure. We need to handle all kinds of data inputs on the one hand, and to describe all analyzing influencing environment entities on the second hand. For example, if we would like to analyze a data set with regression techniques, then we need the number of points, their *x* and *y* coordinates, the number of performed regressions (linear, quadratic, exponential) etc. Having summarized, the random correlation framework parameters are:

- *k*, which is the number of data columns;
- *n*, which is the number of data rows;
- *r*, which is the range of the possible numeric values;
- *t*, which is the number of methods.

To describe all structure, matrix form is chosen. Therefore, parameter *k*, which is the number of data rows [also the columns of the matrix], and *n*, which is the number of data items contained in the given data row [also the rows of the matrix], are the first two parameters. The third parameter, range *r* means the possible values, which the measured items can take. To store these possibilities, the lower (*l*) and upper (*u*) bounds must only be stored. For example, *r*(*1*, *5*) means the lower limit is *1*, the upper limit is *5* and the possible values, which we can measure, are *1*, *2*, *3*, *4* and *5*. Range *r* is not a very strict condition because the measured values intervals can be defined many times. A trivial way to find these limits, when *l* is the lowest measured value and *u* is the highest one. They can be sought non-directly as well. These bounds are determined by an expert in this case. E.g., a tree grows every year, but it is impossible to grow 100 meters from practical point of view. The longitude line is infinity, but it is possible to define *l* and *u*. Although in our work integers are used, it is possible to extend this notation for real numbers since the possible continuous nature of the measured data. The continuous form can be approximated with discrete values. In this case, the desired precision related to *r* can be reached with the defined number of decimals. The sign r(*1*, *5*,) means the borders are the same as before, but this range contains all possible values between 1 and 5 up two decimals.

Parameter *t* is the number of methods. We assume that if we execute more and more methods, the RC probability increase. For example, if *t = 3*, that means *3* different methods are performed after each other to find a correlation. This *t = 3* could be interpreted several ways related to the specific random analyzing process. We have the following four Interpretations for *t*:

- *Interpretation 1*: The number of methods;
- *Interpretation 2*: The input parameter's range of the given method;
- *Interpretation 3*: The decision making according to divisional entity level (output parameter);
- *Interpretation 4*: The outlier analysis.

Interpretation 1 is trivial since it represents the number of performed methods. *Interpretation 2* means that an input parameter can influence the results. In this case, the more input parameters' values are used during the analysis, the higher the possibility that true results born. For example, in statistics, the significance level (α) can be chosen by the user. However, different levels of α have a precise statistic background, it is possible to increase this level by the scientists, which cause H_0 true sooner or later. In other words, if we have more and more data rows, we have connections between them at higher probability. But the Type I error is in the background, which increases in the case of more data rows. Type I error means that we reject

H_0, however it is true. Extending his theory, we state that if we use the correction of Bonferroni, then we still can have random results.

Interpretation 3 is similar to the second one, but this regards to output parameters. It is such an entity, which value can influence the decision. While *Interpretation 2* refers to a calculated number, *Interpretation 3* means that entity, which will be compared with the calculated value. For example, in the case of regressions, choosing r^2 level can be different. There are rules to define which result is "correlated" or "non-correlated". However, these rules are not common and there can be agreed that results with *0.8* or *0.9* are "correlated", but *0.5* or *0.6* are not so trivial. This kind of output divisional entities values have strongly effect on decisions.

Interpretation 4 is necessary most of the cases, however, by performing more and more outlier analysis, the *t* can increase RC factor heavily. It is trivial, that the junk data must be filtered. However, the main problem is still that if we perform more and more techniques then we will get mathematically proved but random result. Moreover, by combining *Interpretations 1-4*, we get a result anyway, decreasing the chance of "non-correlated". For example, combining regression techniques with outlier analysis, the "no good" points filter possibilities can raise and the result gets seemingly better. However, the result has no choice but becoming "correlated", which can lead us to RC.

Models

In this section, methods are introduced, with which the RC occurrence possibility can be calculated. The models can be combined, the given analysis can be examined by all models. There are three main models:

- Ω-model: We calculate the total probability space;
- Θ-method: We determine the chance of getting a collision e.g., find a correlation.
- Γ-model: We analyze that the subsets of the data pass the given method of analysis or not.

In the case of the Ω-model, all possible measurable combinations are produced. In other words, all possible n-tuples related to $r(l,u)$ are calculated. Because of parameter *r*, we have a finite part of the number line, therefore this calculation can be performed. That is why *r* is necessary in our framework. All possible combinations must be produced, which the researchers could measure during the data collection at all. After producing all tuples, the method of analysis is performed for each tuple. If "correlated" judgment occurs for the given setup, then we increase the count of this "correlated" set S_1 by 1. After performing all possible iterations, the rate *R* can be calculated by dividing S_1 with the cardinality of the probability space ($|\Omega|$). *R* can be considered as a measurement of the "random occurring" probability related to RC parameters. In other words, if *R* is high, then the possibility of finding a correlation is high with the given method and with the related *k*, *n*, *r* and *t*. For example, if *R* is *0.99*, the "non-correlated" judgment can be observed only 1% of the possible combinations. Therefore, finding a correlation has a very high probability. Contrarily, if *R* is low, e.g., *0.1*, then the probability of finding a connection between variables is low.

In the case of the Θ-model, rate *C* is calculated. *C* shows how much data are needed to find a correlation with high possibility. Researchers usually have a hypothesis and then they are trying to proof their theory with data. If one hypothesis is rejected, scientists try another one. In practice, we have a data row A and if this data row does not correlate with another, then more data rows are used to get connection related to A. The question is how many data rows are needed to surly find a correlation. We seek that number of data rows, after which correlation will be found with high probability. There is a rule of thumb stating that from 2 in 10 variables (as data rows) correlate at high probability level, but we cannot find any proof, it rather is a statement based on experiences. The calculation process can be different depending on the given method of analysis and RC parameters. During the Θ-model calculation process, we generate all possible candidates ($|\Omega|$) based on RC parameters first. Then, we create individual subsets. It is true for each subset that every candidate in the given subset is correlated with each other. We generate candidates after each other and during in one iteration we compare the

current generated candidates with all subsets' all candidates. If we find a correlation between the current candidate and either of the candidates, then the current candidate goes to the proper subset. Otherwise, a new subset is created with one element, i.e., with the current candidate. *C* is the number of subsets. *C* shows us that how many datasets must be measured during the research to surly get a correlation with at least two datasets. Based on value *C*, we have three possible judgements:

- *C* is high. Based on the given RC parameters, it must be lots of dataset to get a correlation with high probability. This is the best result, because the chance of RC is low.
- *C* is fair. The RC impact factor is medium.
- *C* is low. The worst case. Relatively few datasets can produce good correlation.

Using Γ-model, all subsets of the given data items are produced. For all subsets, the given method of analysis is performed. In this case, rate *S* show us that how many subsets eventuate "correlated" result compare to that subsets which do not.

Analyzing the Process

Random Correlations tell that it is feasible to get the mathematically proven results so, that another result could not come out at higher probability, just the given one. If we would like to perform a research starting from data management ending with results and publication, then the given process can (should) be analyzed from RC point of view as well. In the RC framework, to perform an RC analyzing session related to the given analyzing procedure, the following steps were defined.

- Introducing the basic mathematical background of the method of analysis;
- Define what is exactly understood under the method's Random Correlations;
- Introducing which RC parameters are used to analyze the result's random factor;
- Calculation and proving;
- Validation and interpretation.

With the standard RC analyzing session and the suitable RC entities (parameters, models), we can examine the given analysis from the RC point of view. In the next chapter, the statistical method called Analysis of Variance (ANOVA) is examined as an example how the Random Correlations Framework works in practice.

Analysis of Variance

One research including data and method of analyses can be analyzed by different RC point of view with combining different parameters with different RC models and methods. Therefore, RC analyzing session can be very various. In this chapter, we demonstrate a given RC analyzing session in the matter of Analysis of variance (ANOVA) focusing on Ω-model.

Mathematical background

The ANOVA method is used to determine whether the groups' averages are different or not. It is applied widely in different scientific fields. The null hypothesis H_0 states that the averages are equal statistically and the alternative hypothesis H_1 declines the equality statistically.

Since we have *k* groups and in each group, there are *n* values, therefore we have degrees of freedom df_1 and df_2. The first regards the number of groups, therefore $df_1 = k - 1$. The second is related to individual group values, in each group, the *df* is $n-1$, we have *k* groups, so $df_2 = k*(n-1)$. The *F* test statistic follows the Fisher distribution with df_1 and df_2. Therefore, the given *critical value* can be sought in Fisher table with $F_{(df1,df2)}$. The calculation process is summarized in Table 1 [22].

The following expressions were used:

$$SSB = \sum_{j=1}^{k} n_j * (\bar{x}_j - x)^2$$

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - x_j)^2 ,$$

where *k* stands for the number of columns, *n* is the number of rows, *F* is the test statistic. If the *F* test statistic is smaller than the $F_{(df_1, df_2)}$ critical value, then we can conclude that the sample means are equal at the significance level α. If *F* is bigger, then the means are not equal.

Table 1. ANOVA basic calculation process

Differences	Sum of squares	Degrees of freedom	Mean square	F value
Between groups	SSB	$k-1$	MSB	F=MSB/MSW
Within groups	SSW	$k*(n-1)$	MSW	
Total	SST = SSB+SSW			

ANOVA's Random Correlation

As for the Random Correlation, ANOVA has a specific place: both H_0 and H_1 can be meaningful. The "non-correlated" can be defined as the means are statistically similar [H_0], therefore the influencing variable has no effect on the subject ["non-correlated"]. The H_1 means that variable has influence ["correlated"]. In this case, the random correlation means that the H_0 or the H_1 can take priority over against the other. The seeking rate should be around *0.5*. Small deviation is allowed, but huge distortion is dangerous.

The following parameters are used in this case: (1) Number of measurement (*k*); (2) Number of data item related to one measurement (*n*); (3) Range (*r*) where *r* contains the lower bound *l* and upper bound *u*. Parameter *t* is not used, because we have only one method, i.e., ANOVA itself. The Random Correlation method is the Ω-model. It means that we calculate all possible input values according to *k*, *n*, and *r*. For example, if we have *3* groups and we have *5* values in each group, then we have an ANOVA input matrix with size $3*5$. If we assume that the minimum possible measured value is *1*, the maximum possible measured value is *3*, then we need to generate all possible matrix starting with the case when all value is *1* and ending with the case when all values is *3* in the matrix. The number of all possible input matrixes is $3^{15}=14348907$. It is true that, when ANOVA is constructed long ago, the precise mathematical background allows bigger error rate in the case of bigger sample size. However, for a human brain, it is a heavy task to calculate all possible input matrix, therefore numerical calculations without computers is almost impossible. Hence, a computer program was developed using the equations in Table 1 and some own code to determine the seeking *R* rate. The program generates an input matrix, then it performs ANOVA with the given input matrix, and then the program analyzes whether H_0 or H_1 is accepted. It is necessary to number only H_0 acceptances, because divide H_0 acceptances with the number of all possible input matrixes signed with |Ω|, we get the seeking rate *R*. However, the use of ANOVA has assumptions.

ANOVA Assumptions Calculation

The main calculation process of the Ω-model is to produce all possible input matrixes. However, ANOVA must not be performed on input matrixes which are not pass ANOVA assumptions. Since we must follow the precise research methodology, if we measured such "assumption-not-passed" input matrix in the case of a given research, then we cannot use ANOVA as a method of analysis. This kind of input matrix breaks the condition of ANOVA use. Therefore, the ANOVA assumptions must be handled in the implementation. There are three assumption of ANOVA adaptation:

- Sampling must be done randomly;
- Each group must follow the normal distribution (normality check);
- Variances must be statistically equal (homogeneity check).

We can assume that the first assumption is passed. The combinations, which do not follow the normal distribution must be deleted from the set of candidates, i.e., from the set of input matrixes. We used the D'Agostino-Pearson test to check normality. Another assumption is that the variances must be statistically equal. The candidates which variances are not equal must also be deleted from the set of

input matrixes. The equality of variances was checked with Bartlett test. After handling ANOVA assumptions, *R* can be calculated based on the rest of the input matrixes which pass all assumptions. The whole calculation process is summarized in Figure 12. This process is suited for the calculation of the seeking *R*. However, if the RC parameters' values are increasing, then the |Ω| is exponentially increasing. Therefore, the calculation process cannot carry out even with a computer. The fact that some matrixes do not pass the assumptions does not help, because all matrixes must be checked, i.e., all matrixes must be produced once. It does not matter that the given matrix passes the assumptions or not from the algorithm cost point of view. Because of the exponential growth of |Ω|, dimension reduction must be applied to calculate *R* in the case of higher values of RC parameters.

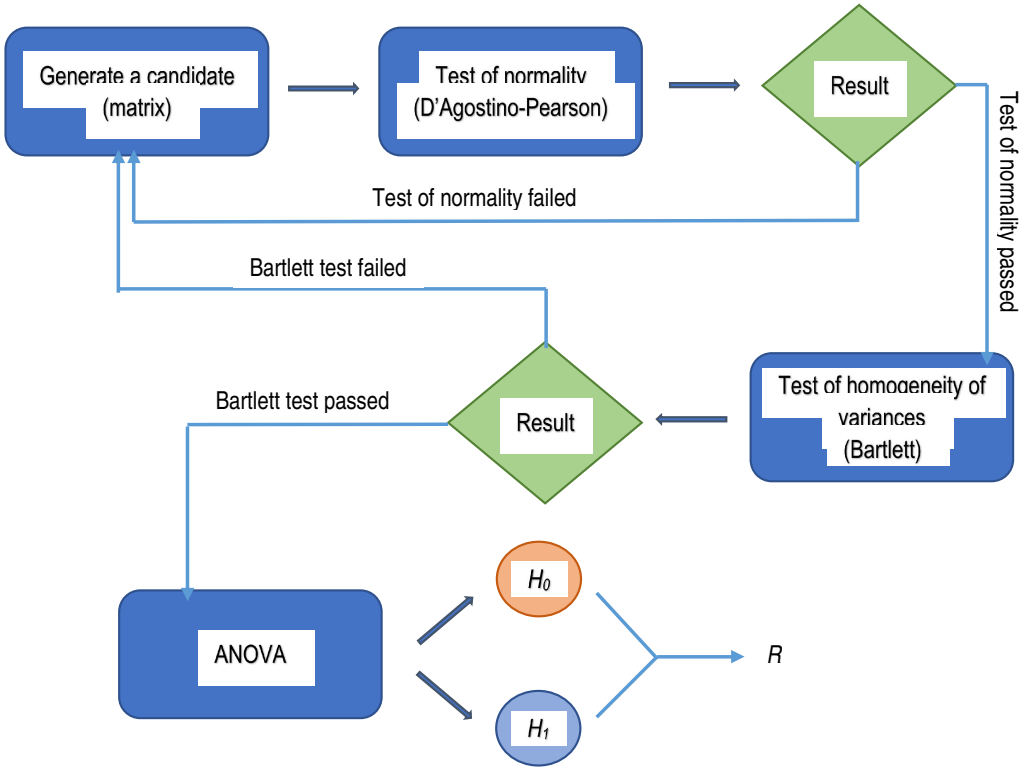


Figure 12: Calculation process of R in the case of ANOVA

Dimension Reduction and Results

With the basic calculation process, we can analyze ANOVA random property only in the case of low values of RC parameters. To solve the |Ω| exponential growth problem, we can utilize that the (*a*) and (*b*) matrixes have same *SSB*, *SSW* and degrees of freedoms, the *MSB* and *MSW* are also the same including *F* value.

$$\begin{bmatrix}1 & 2 & 1\\ 2 & 3 & 3\\ 3 & 2 & 1\end{bmatrix}(a) \quad \begin{bmatrix}3 & 3 & 3\\ 2 & 2 & 1\\ 2 & 1 & 1\end{bmatrix}(b)$$

Based on that property of ANOVA, we developed a new algorithm named as Finding Unique Sequences (FUS). Now, we can calculate the seeking rate *R* with bigger RC parameters' values. This algorithm is introduced in detail in our previous paper [23]. FUS can moderate the calculation hardness of |Ω|. With FUS, ANOVA can be analyzed from the RC point of view in the case of higher values of RC parameters.

However, also FUS has a limitation and further increasing values cause such big |Ω|, which cannot be calculated relatively short time even with FUS either. To solve this problem, simulation technique is applied. The precise description of the simulation process is described in our previous paper [23]. We mark the simulation result of *R* with *R'*. The FUS algorithm and the simulation technique with ANOVA implementation, assumptions and Ω-model are developed into a computer program written in C# language. All parts of the program such as ANOVA implementation, candidate generations, D'Agostino-Pearson and Bartlett test are validated by other trusted software such as Excel and R statistical software environment.

Table 2: ANOVA results related to *R*

r(1,3)	*k* = *3*	*k* = 4	*k* = 5	*k* = 6	*k* = 7	*k* = 8
n = 4	1	0.999	0.999	0.999	0.999	0.999
n = 5	0.987	0.992	0.994	0.995	0.996	0.997
n = 6	0.974	0.978	0.981	0.984	0.986	0.987
n = 7	0.969	0.974	0.976	0.979	0.981	0.983
n = 8	0.963	0.966	0.968	0.970	0.972	0.973
n = 9	0.956	0.959	0.961	0.962	0.963	0.964
n = 10	0.965	0.969	0.972	0.974	0.975	0.976

During the calculations, we use significant level α = 0.1 in all cases. The first result of *R* can be seen in Table 2. In the case of small *n*, *k*, and *r*, the rates show high probability in the favor of H_0. If the values of RC parameters are small, then whatever researchers measure, the result will be H_0 at very high probability, the compared data are statistically equal on average. The specific case is *n = 4* and *k = 3*, because the chance of H_0 is *1*, which means there is no such input matrix eventuating H_1. The parameter values are changed in the further calculations and FUS algorithm must be applied in the case of increased parameters' values. Using FUS, the results with different *r*, *k*, and *n* are shown in Table 3. We can conclude that H_0 is still dominant, however, *R* values slightly decrease in the case of higher values of RC parameters.

Table 3: *R* alues using FUS

r(*l*, *u*)	*k*	*n*	*R*
(1, 5)	3	4	0.975
(1, 5)	3	5	0.958
(1, 5)	3	10	0.977
(1, 5)	4	5	0.958
(1, 5)	4	9	0.959
(1, 5)	5	6	0.959
(1, 10)	3	4	0.954
(1, 10)	3	5	0.957
(1, 10)	4	4	0.955

According to Big Data inspired environment, larger *n*, *k* and *r* should be used. Therefore, the simulation technique must be applied. The results of simulations are summarized in Table 4. This table has two parts. In the first part, *R* and *R'* can be compared. The results show that approximation of *R'* is appropriate. In the second part, such cases are shown in which *R* cannot be calculated with FUS either. In these cases, only *R'* can be calculated in real time. First, the rates are high in the favor of H_0. But in the case of large enough *k* and *n*, the rates are heavily turn to H_1. If the same experiment is performed with relatively small RC values, getting the result H_0 and the "non-correlated" judgment is very high. Contrarily, the chance of H_1 is increased with large enough values and "correlated" decision will be accepted at high probability. However, this is a paradox in the view of big data inspired environment. In general, if we have a conclusion with smaller number of data items, then more data should enhance the conclusion.

ANOVA can be analyzed not just with Ω-model. From the Θ-model point of view, rate *C* shows that number, which is needed to surly accept H_1 near the given values of RC parameters. In other words, higher number than *C* causes such environment, near which one column, i.e., one measurement, surly differs from the others statistically at the given significant level. In the ANOVA case, the Γ-model means that we not just simply increase parameter *n*, but we analyze the measured data so, that we skip rows systematically. For example, if we have *30* rows, i.e., *n = 30*, then we should analyze

$$n-1\ (2_9),\ n-2\ (2_8),\ n-3\ (2_7),$$

etc. input matrixes. In each iteration, the decision of ANOVA is registered. If the judgments change very often, i.e., rate *S* is high, it means that the judgment in the case of *n = 30* is not very stable. In other words, if *S* is high, then the probability of changing the judgment in the case of *n = 31*, or *n = 32* is high.

Table 4: ANOVA results of *R* and *R'*

r(l, u)	k	n	R	R'	r(l, u)	k	n	R'
(1,3)	3	30	0.952	0.934	(1,3)	4	100	0.915
(1,3)	3	50	0.954	0.973	(1,3)	7	100	1.09E-9
(1,3)	5	10	0.972	0.962	(1,3)	10	100	0
(1,3)	5	15	0.960	0.989	(1,3)	10	500	0
(1,5)	3	10	0.977	0.924	(1,5)	4	100	0.588
(1,5)	4	5	0.958	0.978	(1,5)	5	100	0.004
(1,5)	4	9	0.959	0.953	(1,5)	7	100	7.2E-19
(1,10	3	5	0.957	0.943	(1,10)	4	10	0.971
(1,10	4	5	0.956	0.967	(1,10)	4	19	0.960

Conclusion

It is important to remark that we do not deny that real connections exist. We state that Random Correlations should be considered in the cases of all researches. Therefore, the standard research methodology steps, such as research design, data collection rules, analysis execution, result interpretation, should be extended with the step of RC analysis. It means that scientists may calculate RC factor based on data and given method of analysis. To distinguish real correlations from Random Correlations, we recommend for all scientists to always calculate how big the probability of RC can be. It is important to analyze whether the results space balanced or not. It is true that there are models and methods such as Bootstrap, Cramer's Coefficient or Correlation Coefficient, with which the RC theory must be compared. However, the RC factor should be attached to every scientific result.

References

[1] Khan, J.A. (2008): Research methodology, APH Publishing Corporation, New Delphi, 2008, pp334, ISBN: 8131301362

[2] Kuada, J. (2012): Research Methodology: A Project Guide for University Students, Samfundslitteratur, Frederiksberg, 2012, pp139, ISBN: 8759315547

[3] Lake, P., Benestad, H. B., Olsen B. R. (2007): Research Methodology in the Medical and Biological Sciences, Academic Press, London, 2007, pp519, ISBN: 0123738741

[4] Mohapatra, A., Mohapatra, P. (2014): Research methodology," Partridge Publishing, India, 2014, pp124, ISBN: 148281790X

[5] Zavaleta, E. S., Thomas, B. D., Chiariello, N. R., Asner, G. P., Shaw, M. R., Field, B. C. (2003): Plants reverse warming effect on ecosystem water balance, Proceedings of the National Academy of Sciences of the United States of America, 2003/100, pp9892–9893

[6] Liu, W., Zhang, Z., Wan, S. (2009): Predominant role of water in regulating soil and microbial respiration and their responses to climate change in a semiarid grassland, Global Change Biology, 2009/15, pp184–195

[7] Church, J. A., White, N. J. (2006): A 20th century acceleration in global sea-level rise, Geophysical Research Letters, 2006/33, pp1–4

[8] Houston, J.R., Dean, R.G. (2011): Sea-Level Acceleration Based on U.S. Tide Gauges and Extensions of Previous Global-Gauge Analyses, Journal of Coastal Research, 2011/27, pp409–417

[9] Hooper, L., Bartlett, C., Smith, G. D., Ebrahim, S. (2002): Systematic review of long term effects of advice to reduce dietary salt in adults, British Medical Journal, 2002/325, pp628–632

[10] Pljesa, S. (2003): The impact of Hypertension in Progression of Chronic Renal Failure, Bantao Journal, 2003/1, pp71-75

[11] Held, I. M., Delworth, T. L., Lu, J., Findell, K. L., Knutson, T. R. (2006): Simulation of Sahel drought

in the 20th and 21st centuries, Proceedings of the National Academy of Sciences of the United States of America, 2006/103, pp1152–1153
[12] Haarsma, R. J., Selten, F. M., Weber, S. L., Kliphuis, M. (2005): Sahel rainfall variability and response to greenhouse warming, Geophysical Research Letters, 2005/32, pp1–4
[13] Giannini, A. (2010): Mechanisms of Climate Change in the Semiarid African Sahel: The Local View, Journal of Climate, 2010/23, pp743–756
[14] Ng, B., Wiemer-Hastings, P. (2005): Addiction to the Internet and Online Gaming, Cyberpsychology & Behavior, 2005/8, pp110–113
[15] Yee, N. (2006): Motivations for Play in Online Games, Cyberpsychology & Behavior, 2006/9, pp772–775
[16] Shindell, D. T., Miller, R. L., Schmidt, G. A., Pandolfo, L. (1999): Simulation of recent northern winter climate trends by greenhouse-gas forcing, Nature, 1999/399, pp452–455
[17] Petoukhov, V., Semenov, V. A. (2010): A link between reduced Barents-Kara sea ice and cold winter extremes over northern continents, Journal of Geophysical Research, 2010/115, D21111
[18] Knippertz, P., Ulbrich, U., Speth, P. (2000): Changing cyclones and surface wind speeds over the North Atlantic and Europe in a transient GHG experiment, Climate Research, 2000/15, pp109–122
[19] Vautard, R., Cattiaux, J., Yiou, P., Thépaut, J.-N., Ciais, P. (2010): Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness, Nature Geoscience, 2010/3, pp756–761
[20] Bogardi, I., Matyasovzky, I. (1996): Estimating daily wind speed under climate change, Solar Energy, 1996/57, pp239–248
[21] Nosek, B. et al. (2015): Estimating the reproducibility of psychological science, Open Science Collaboration, Science, 2015/28, aac4716
[22] Sahai, H., Ageel, M. I. (2000): Analysis of variance: fixed, random and mixed models, Springer, pp742,
[23] Bencsik, G., Bacsárdi, L. (2016): Novel methods for analyzing random effects on ANOVA and regression techniques, Advances in Intelligent Systems and Computing, Springer, pp499-509

Forecasting Performance Improvement: Comparison of Auto-mated Statistical and Neural-Network Based Methods

TIBOR KOVÁCS
Corvinus University of Budapest, assistant lecturer
eMail: tibor.kovacs@uni-corvinus.hu

ABSTRACT

The aim of the paper is to evaluate the applicability of automated forecasting methods for a specific business problem: estimating future performance improvements. The study is using data sourced from a global consumer goods company and evaluates forecasts for three performance indicators for 63 sites. Time series decomposition (STL), exponential smoothing (Holt-Winters), autoregressive moving average (ARIMA) and multilayer perceptron artificial neural-networks based forecasting methods have been used with automated model selection and training. The models' performance have been compared based on their demonstrated accuracy. The study highlights that automated forecasting could be applied for business problems that may fall outside of typical use of these methods, even when the data sets do not follow a uniform pattern. However, the findings also indicate that these forecasts might be inaccurate. The limitations of this study are that only five methods have been evaluated, and model parameters have only been fine-tuned to a limited extent.

Introduction

Forecasting is a common business task: it is widely used to predict sales demand, schedule production or commit raw material purchases. Forecasting performance improvements and quantifying their value is an area that is studied to a lesser extent, although it could bring significant business benefits. Financial markets demand publicly listed companies to predict accurately their future financial position, where forecasting the added value derived from performance improvements could play an important role. This research evaluates the applicability of different forecasting techniques for this specific business problem. The techniques belong to two distinct groups: The Seasonal and Trend decomposition using Loess (STL), the exponential smoothing (Holt-Winters) and the Autoregressive Integrated Moving Average (ARIMA) methods are statistical techniques, while the feedforward Multilayer Perceptron (MLP) Artificial Neural Networks (ANN) are soft computing or computational intelligence techniques. While seasonal and trend decomposition is primarily used to explore and study the patterns time series exhibit, it could also be used for forecasting [13], hence the inclusion of the STL method.

Time series forecasting requires having numerical information available about the past and the assumption, that some aspects of past patterns will continue into the future. If the system is well understood and the predictor variables are all available, explanatory models could be constructed, describing the exact relationship between the predictor variables and the outcome. Various forms of linear regressions could be the examples for these methods. However, there are limitations of using explanatory models: the systems are often complex, they may not be well understood and the relationship between the output and predictor variables may be difficult to measure. Furthermore, knowing the future values of certain predictor variables may be difficult or impossible; therefore, time series forecasting may give more accurate results than explanatory models [13]. This study is using performance data sourced from a global consumer goods company, SABMiller plc, that was facing the challenges of forecasting and monetizing the added value derived from a lean capability development program. The program resulted to the improvement of several performance indicators over multiple sites and the financial benefits had to be estimated to accurately project the company's future financial positions. Forecasting in real business environment often presents the challenges of having to process numerous time series simultaneously, that may have substantial noise and may not follow uniform patterns or seasonal profiles. Automated forecasting methods, where the forecast parameters are calculated without human intervention could be a solution for these types of business problems. The research evaluates the performance of automated forecasts, that are developed for three performance indicators over 63 business units, for a one-year ahead horizon. The performance of the methods is evaluated based on their forecasts accuracy.

Decomposition using Loess

When exploring the patterns of time series data, it is typically split into three components: a trend component, a seasonal component, and a remainder component (that could also be described as noise). The decomposition could be additive or multiplicative:

$$Y[t] = T[t] + S[t] + e[t] \text{ (additive)} \quad (1)$$

$$Y[t] = T[t] * S[t] * e[t] \text{ (multiplicative)} \quad (2)$$

where Y[t] is the time series data, decomposed to the trend T[t], the seasonal S[t], and the reminder components e[t]. Multiplicative decomposition is more appropriate, when the variation in the seasonal pattern appears to be proportional to the level of the time series [13]. The STL decomposition method, used in this research was developed by Cleveland, Cleveland, McRae and Terpenning [4]. It is utilising local regression (Loess) for smoothing and decomposing the time series using two loops: In the outer loop, robustness weights are assigned to each data point depending on the size of the remainder. This allows for reducing or eliminating the effects of outliers. The inner loop iteratively updates the trend and seasonal components, by subtracting the current estimate of the trend from

the raw series. The time series is then split into cycle-subseries, that are loess smoothed and then passed through a low-pass filter. The seasonal components are the smoothed cycle-subseries minus the result from the low-pass filter. The trend component is the seasonal components subtracted from the raw data, that is loess smoothed. The remainder component is the trend and seasonal components subtracted from the raw series [7].

When decompositions are used for developing forecasts, it usually assumes, that the seasonal component is unchanging, therefore they apply last year's seasonal pattern for the forecasts. The seasonally adjusted component (trend and noise) can be forecast using different methods. In this research this is done by using exponential smoothing, using automatic model parameter estimation [12].

The Holt-Winters' seasonal forecast

Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations get older. This framework generates reliable forecasts quickly and for a wide range of time series. The Holt-Winters seasonal method [9, 10, 24] comprises three smoothing equations — one for the level ℓ_t, one for the trend b_t, and one for the seasonal component s_t, with corresponding smoothing parameters α, β* and γ. (m denotes the frequency of the seasonality, e.g. 12 for monthly data). The equations for the multiplicative model (as this is more appropriate for the data of this research) are:

$$\hat{y}_{t+h} = \ell_t + hb_t + s_{t+h-m(k+1)} \quad (3)$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \quad (4)$$

$$s_t = \gamma\left(\frac{y_t}{\ell_{t-1} + b_{t-1}}\right) + (1-\gamma)s_{t-m} \quad (5)$$

where ($\hat{y}_{t+h}$) is the forecast value for *t+h* time, *k* is the integer part of *(h-1)/m* which ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample [13].

The model parameters are estimated using the maximum likelihood method.

The ARIMA model

One of the most prominent techniques to model time series are the Autoregressive Integrated Moving Average (ARIMA) class of models. These models are often referred to as Box- Jenkins models, because of the work of these two researchers [1, 2]. These models are data driven: rather than a priori specifying the components of the structural model such as level, trend, and seasonality, they are using statistics to examine the data and select the model that fits it best. These class of models are relying on the concept of stationarity: a stationary time series is one whose properties do not depend on the time at which the series is observed. Time series with trends, or with seasonality, are not stationary, the trend and seasonality will affect the value of the time series at different times. A white noise time series, that has a zero mean and σ^2 variance, on the other hand is stationary, it looks much the same at any point in time [13]. Non-stationary time series could be transformed to a stationary one by differencing it one or more times, or for seasonal time series differencing them to the number of seasons between similar observations. The ARIMA class of models could be decomposed to autoregressive (AR), integrated (I) and moving average (MA) components. The AR model forecasts the variable as the linear combination of past values of the same variable, hence the term autoregression (regression against itself). The AR component of an order *p* can be written as (6):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where y_t , y_{t-1}, y_{t-p} are the variable at time *t, t-1, t-p;* Φ_1, Φ_2, Φ_p, are the model parameters and ε_t is the random error term or white noise. The parameter *c* may be interpreted as an intercept term, whereas *Φ* is a slope term. The MA component of an order *q* can be written as (7)

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

which is like the AR model except instead of using past values of the variable, it is using past forecast

errors in the regression like model. yt is the variable at time t, ε_t is the white noise, ε_{t-1}, ε_{t-2}, ε_{t-p} are forecast errors at time t, t-1, t-p and θ_1, θ_2, θ_q are the model parameters. When the time series has a trend or a seasonal pattern, differencing can help to stabilise the mean of the time series and therefore eliminating the trend or the seasonality, resulting to a stationary time series of white noise. A time series with drift for example could be first order differenced resulting to white noise, that can be written as

$$y_t = c + y_{t-1} + \varepsilon_t \quad (8)$$

where c is the average of the changes between consecutive observations that is called drift. Seasonal data requires to be differenced to the number of seasons between similar observations where m is the number of seasons, also called "lag-m differences"

$$y_t = y_{t-m} + \varepsilon_t \quad (9)$$

The non-seasonal ARIMA model is obtained by combining the autoregression, first order differencing and moving average models. This model can be written as (10)

$$y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p}' + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-q} + \varepsilon_t$$

where s_t is the differenced time series and the predictors are both the lagged vales and the lagged errors. This model is called ARIMA(p, d, q) model, where p is the order of the autoregressive part, d is the degree of differencing and q is the order of the moving average part. In case of seasonal time series, additional seasonal terms are included in the model and called as ARIMA(p, d, q)(P,D,Q)m where m is number of observations per year, the uppercase notation are for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model.

There are several methods to find the appropriate values of (p, d, q)(P,D,Q) and m as described by [15]. The "forecast" R package [14, 15] applied in this research is using KPSS unit-root test [18] for non-seasonal data and Canova-Hansen test [3] or an estimation of seasonal strength [23] to finding m, with subsequent KPSS unit-root test to select automatically the orders of the model. Finally, model parameters are then estimated using maximum likelihood estimation, assuming that the white noise components are normally distributed.

The MLP-ANN model

Artificial neural networks (ANN) are the other well-known techniques in time series forecasting due to their proven capabilities in approximating any linear or nonlinear functions [11]. Most commonly, they use a feed-forward topology of the multi-layer perceptron (MLP) with single output node (Figure 1-1), where a single h-step ahead forecast

$$(\hat{y}_{t+h})$$

$$(\hat{y}_{t+h})$$

is calculated using the input vector of *p+1* lagged observations

$$(y_t, y_{t-1}, \cdots, y_{t-p})$$

.

In case of multistep-ahead forecasting, one way of implementing it using a single output node ANN is the so-called iterative process, when forecast values are used iteratively as inputs for the next forecasts.

The forecasting performance of the MLP-ANN model is dependent on the number of input nodes, hidden layers and hidden nodes. Finding the appropriate combination of these parameters is not trivial and in most cases problem-dependent. In general, networks with fewer hidden nodes are preferable, as they usually have better generalization ability and having overfitting problem to a lesser extent. On the other hand, networks with too few hidden nodes may not have enough power to model and learn the data [8], hence the optimal combination of hidden layer and nodes needs to be found. The input vector is not necessarily constructed from consecutive lagged observations, those dynamic lags has to be selected for optimal performance, that capture the components of level, trend and seasonality, while remaining robust against outliers and noise [5]. The "nnfor" R package [17] of MLP-ANN time series forecasting, applied in this research is using several methods to automatically select the most appropriate combination of lagged input variables. The method tests the time series for

non-stationarity (i.e. trend or structural breaks through level shifts) using the augmented Dickey–-Fuller (ADF) test. For non-stationary time series candidate input vectors of lagged observations of the original time series and the detrended time series are created using (first or second order) differencing. For seasonal time series candidate input vectors of lagged observations of the original time series and various other combinations are considered including de-trended and de-seasonalised lagged observations (using seasonal differences). The MLP-ANN is then trained on these candidates and the best performer of the candidate input vectors of lagged observations is selected for the forecast model.

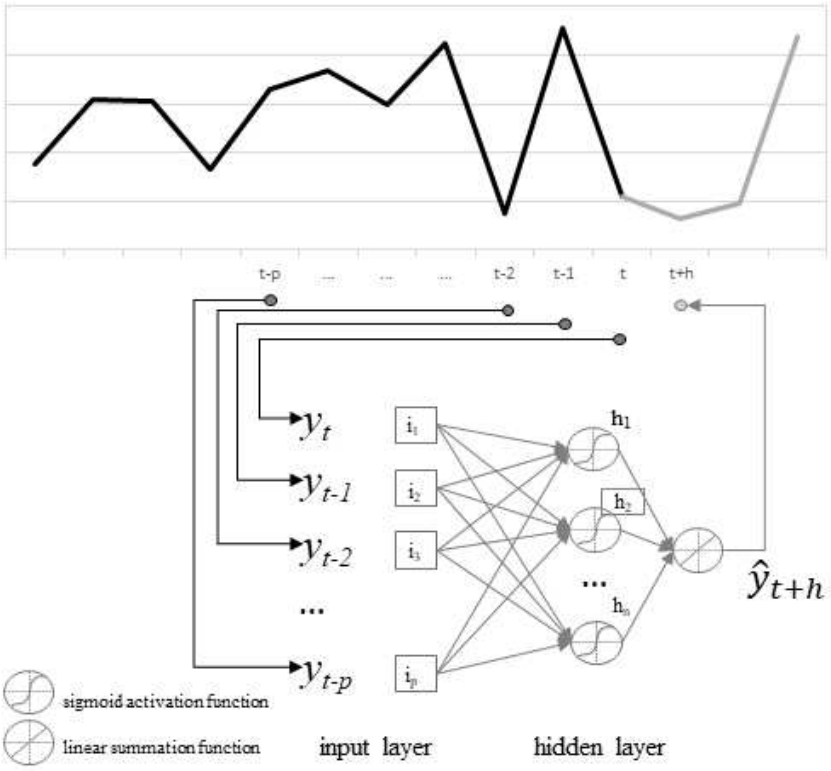


Figure 1-1. MLP-ANN architecture for time series forecasting (based on Ord et al. [21]

Measuring forecast accuracy

Alternative methods could be compared using forecasting performance measures to select the ones that performs best for a particular problem. These performance measures could also be used, once the method has been selected and being put in use, to monitor, if the method maintains its performance over time. Time series forecast performance measures are based on the forecast error (e_{t+h}), that is the difference between an observed value (y_{t+h}) and its forecast ($\hat{y}_{t+h|t}$). Here "error" does not mean a mistake, it means the unpredictable part of an observation [13]

$$e_{t+t} = y_{t+h} - \hat{y}_{t+h|t} \quad (11)$$

There are various scale-dependent and scale-free performance measure. Mean Absolute Error (MAE) is a scale-dependent measure (12).

$$MAE = \frac{\sum_{h=1}^{m} |y_{t+h} - \hat{y}_{t+h|t}|}{m} = \frac{\sum_{h=1}^{m} |e_{t+h|t}|}{m}$$

Mean Absolute Percentage Error (MAPE) is a scale free measure therefore, it can be used to make comparisons across multiple time series (13).

$$MAPE = \frac{100}{m}\sum_{h=1}^{m}\frac{|y_{t+h} - \hat{y}_{t+h|t}|}{y_{t+h}} = \frac{100}{m}\sum_{h=1}^{m}\frac{|e_{t+h|t}|}{y_{t+h}}$$

Root mean square error (RMSE) gives greater weight to large (absolute) errors. It can be shown that RMSE ≥ MAE for any set of m forecasts [21] (14).

$$RMSE = \sqrt{\frac{\sum_{h=1}^{m}(y_{t+h} - \hat{y}_{t+h|t})^2}{m}} = \sqrt{\frac{\sum_{h=1}^{m}e_{t+h|t}^2}{m}}$$

When evaluating the performance of a forecasting method or comparing several ones, it is important to analyse the underlying error distributions. Although they may indicate the same ranking of relative performance, when comparing different methods, they may be more or less sensitive for outliers, that are present is the data sets [6].

This research is using MAPE to compare the performance of the different models, as it is scale free. Although it has the disadvantage of being infinite or undefined if $y_t=0$ for any t in the period and having an extremely skewed distribution when any value of y_t is close to zero - as all data were positive and greater than zero in this study - MAPE was selected for reasons of simplicity [16].

Research questions

This research is aimed at evaluating the following questions in the context of a specific business problem: quantitatively forecasting performance improvements of multiple performance measures for multiple manufacturing sites.

- How well applicable are the STL, Holt-Winters, ARIMA and/or the MLP-ANN methods for forecasting performance improvements for multiple sites?
- What is the accuracy of these methods?
- What are the limiting factors of using these methods if any?

These questions are answered by developing forecasts for three different performance measures for 63 sites and comparing their accuracy. As the performance measures do not follow a uniform pattern – there are differences in their trend, seasonality and inherent noise – this may reveal insights, that could be used when selecting the right method.

Research methodology

This research is using SABMiller plc as a case study example for this particular business problem. SABMiller plc was a highly successful global beverage company with manufacturing footprint over all five continents. It delivered a total shareholder return of 913.3% between 1999 and 2015 [19], before being taken over in 2016 by its rival, the transaction being the third-largest corporate takeover on record at that time [20]. Continues performance improvement was an essential part of the company's strategy. The company took a systematic approach to improving performance by implementing a lean-practice framework. The expected future improvements were integrated into the company's financial process by monetizing them and including them in the annual budget. Accurate forecasting of these improvements was important for the company: under-forecasting them would have reduced the pressure on management, while over-forecasting would have put the company's financial commitments at risk.

In this research three key performance indicators have been selected to develop forecasts and evaluate their accuracy: extract loss, water usage and energy usage. Extract loss measures the percentage loss of the most significant raw material during the process; water and energy usage measures the amount of water and energy used to produce a unit of finished product. Smaller numbers represent better performance for all three performance indicators. They all constituted significant value, improving these indicators could have delivered several million dollars annually to the group. The performance indicators are sourced from the company's performance database, they were measured monthly by each manufacturing site. Five years (61 months) of data was available for the period June 2011 - June 2016 for 63 sites. There could be several predictor variables, that may influence these performance indicators, however the

relationship between the output and the predictors are too complex to develop an explanatory model. Also, some of the variables cannot be measured accurately or could not be made available as an input variable for forecasting. Two variables, that thought to be important were the monthly production volume and the level of lean practice implementation, the latter is measured through survey instruments, as part of the lean practice implementation process and expressed as maturity score. Certain usage related performance indicators often improve with increased production volume as there could be a fixed usage component that is diluted with higher volume. Higher levels of maturity should lead to better performance, as the plant operates more efficiently. Production volume and lean maturity scores were tested as external regressors to both ARIMA and MLP-ANN forecast models, however as these forecasts were less accurate than the ones that without external regressors, they were discarded. For the forecasts therefore, only lagged observations of the performance indicator were used, without external regressors.

First descriptive statistical analysis of the three performance indicators were done using the "decompose" function of R [22]. A multiplicative decomposition model was used, first determining the trend component as a moving average, using a symmetric window with equal weights and removing it from the time series. Then, the seasonal figure was computed by averaging, for each time unit, over all periods. The seasonal figure was then centred. Finally, the remaining, error component was determined by removing trend and seasonal figure from the original time series. This method of decomposition is different from the one, that was used for forecasting. It was chosen, as a simple method to understand the potential differences between the patterns the three performance indicators exhibit.

The data was then split to training and testing sets; the first 49 months of data were used for training the models, while the last 12 months were used for testing. The STL decomposition for forecasting was performed using the "stl" function of the "stats" R package [20] using the settings setting "periodic" for seasonality and "robust" for handling outliers. The STL forecasts were fitted using the "forecast" function of "forecast" R package on the decomposed time series. This function is applying exponential smoothing on the trend and remaining components of the decomposed time series and then applying last year's seasonality on this forecast. The Holt-Winters and ARIMA forecasts were calculated with the "hw" and "auto.arima" functions of the "forecast" R package [14, 15], for the Holt-Winters forecasts forcing the use of multiplicative models. The ARIMA forecasts were fitted for each performance indicator - site combination, first automatically selecting the most appropriate combination of $(p,d,q)(P,D,Q)$ and m from various candidates and then estimating the model parameters. Separate forecasts were fitted for each site and each performance measure combination and MAPE forecast accuracy measure was calculated for the 12 months testing sets. MLP-ANN forecasts were fitted with the "mlp" function of the "nnfor" R package [17]. Two different networks were trained: the first network "*MLP-ANN fixed lag"* with a fixed set of 7 lagged observations (-1,-2,-3,-5,-7,-9,-11 month lags), with one hidden layer with a maximum of 5 hidden nodes. The second network *"MLP-ANN auto"* has selected the input vector of lagged observations dynamically [5], had also one hidden layer and a maximum 5 hidden nodes.

Results and discussion

The descriptive statistical analysis shows, that all three decomposed performance measures exhibit a trend component that decreases over time, a seasonal component of varying degree but cyclic over 12 months and a remaining, random component. Figure 1-2. illustrates the multiplicative decomposition of one performance indicator (extract loss) for one site (no. 25). The trend and the random components of the decomposed time series have similar patterns, having similar statistical measures. The trend component has similar ranges, suggesting that similar improvements were achieved across all three measures. The seasonal components of the decomposed time series however show some differences; extract loss has the lowest range, while energy usage has the highest range of their seasonal component, suggesting that the latter being

more seasonal than the former. Table 1-1. shows the mean and the standard deviation of ranges, as well as the minima and maxima of the three components, the trend component being normalised by dividing the values by their mean (for each time series), for comparability.

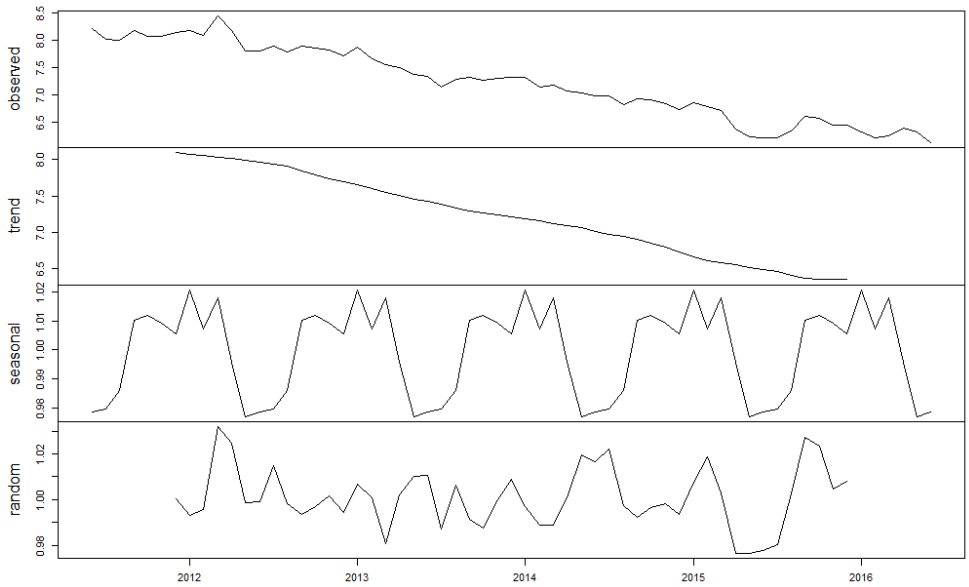


Figure 1-2. Decomposition (multiplicative) of time series - extract loss, site.no. 25

Table 1-1. Statistical analysis of decomposed (multiplicative) time series, trend components are normalised by their mean, 63 sites

		extract loss		water usage		energy usage	
		mean	standard deviation	mean	standard deviation	mean	standard deviation
trend component	range	0.2905	0.1657	0.2711	0.1517	0.2684	0.1424
seasonal component	range	0.1801	0.1024	0.2076	0.2092	0.3088	0.2871
random component	range	0.4130	0.2349	0.3389	0.2630	0.3751	0.2610
trend component	maximum	1.1551	0.1001	1.1563	0.1048	1.1606	0.0992
	minimum	0.8646	0.0751	0.8852	0.0593	0.8923	0.0539
seasonal component	maximum	1.0969	0.0620	1.1285	0.1787	1.1919	0.2294
	minimum	0.9168	0.0480	0.9209	0.0414	0.8831	0.0689
random component	maximum	1.2209	0.1346	1.1925	0.1834	1.2138	0.1869
	minimum	0.8079	0.1346	0.8536	0.0869	0.8387	0.0834

As part of developing the forecasts, they were plotted for all model - performance indicator - site combinations. Figure 1-3. and Figure 1-4. show two examples for the ARIMA and MLP-ANN models. Forecast accuracy measures were calculated and were stored together with the ARIMA model parameters. The forecasts for all three performance indicators had a wide range in accuracy and they were not particularly precise (Figure 1-5. and Table 1-2., Table 1-3. and Table 1-4.). The statistical forecasts (SLT, Holt-Winters and ARIMA) had the best performance across all three indicators, STL with exponential smoothing and ARIMA being the most accurate. These methods could forecast the

majority of data within 25% MAPE accuracy with the occasional outliers within the 50% MAPE range. The average accuracy of these statistical forecasting methods is in the region of 10% MAPE. MLP-ANN models with fixed set of lagged observations generated the least accurate forecasts, the MLP-ANN models with dynamically selected input vectors having somewhat better performance. This demonstrates, that the input vector selection algorithm delivers measurable benefits. There are outliers of inaccurate forecasts in all three models; worst performing forecasts were found among the MLP-ANN models with fixed set of lagged observations. Analysing the individual forecasts shows, that when past patterns not continuing to the future due to disturbances or interventions, poor forecasts are generated by all models. Sudden peaks, high inherent random error, changes in seasonality profile and the lack of a clear trend present difficulties to forecasting to all models. MLP-ANN models could particularly be affected by these features, through the phenomena of over-fitting.

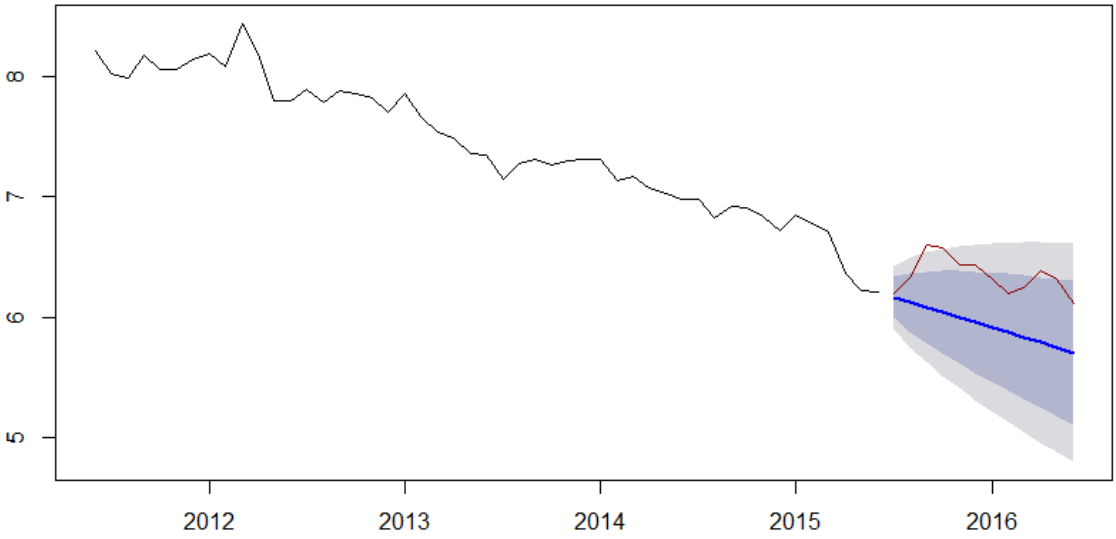


Figure 1-3. ARIMA (auto) forecast for extract loss, site no. 25

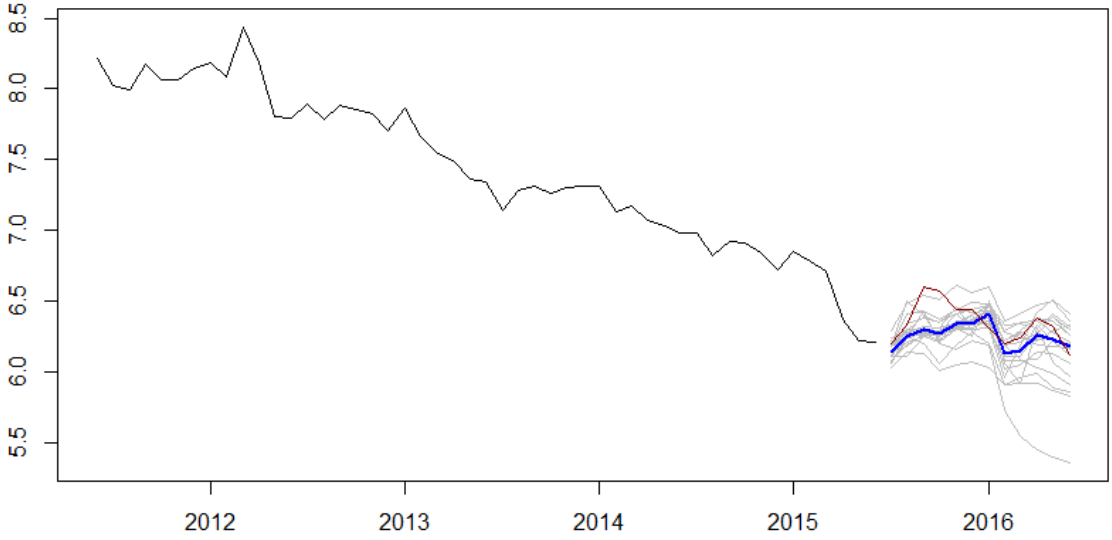


Figure 1-4. MLP-ANN (auto) forecast for extract loss, site no. 25

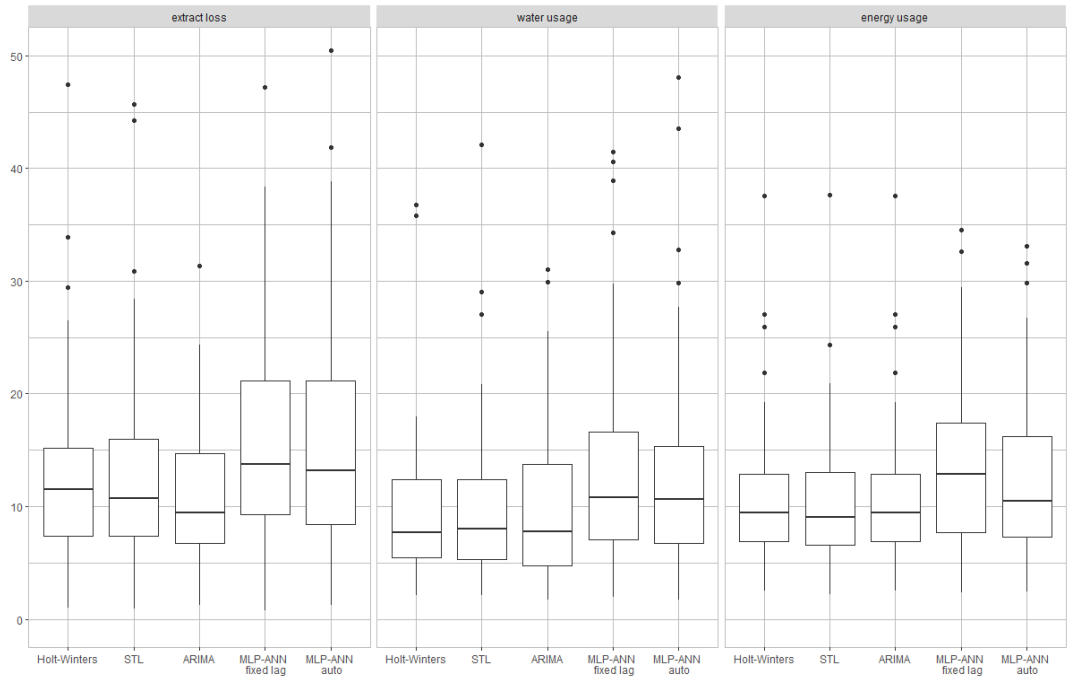


Figure 1-5. Forecast accuracy comparison
(y axis is cut-off at 50, there may be outliers outside of this range. see Table 1-2.)

Table 1-2. Forecast accuracy comparison of different forecasting methods measured as Mean Absolute Percentage Error (MAPE) – extract loss

	extract loss				
	STL	Holt-Winters	ARIMA	MLP-ANN fixed lag	MLP-ANN auto
n	63	63	63	63	63
mean	12.55	12.12	11.09	17.63	15.78
best	0.92	0.98	1.27	0.76	1.23
worst	44.23	47.47	31.33	71.74	52.93
stdev	7.70	7.77	6.31	14.45	11.11

Table 1-3. Forecast accuracy comparison of different forecasting methods measured as Mean Absolute Percentage Error (MAPE) – water usage

	water usage				
	STL	Holt-Winters	ARIMA	MLP-ANN fixed lag	MLP-ANN auto
n	63	63	63	63	63
mean	9.57	10.04	9.89	13.17	12.72
best	2.14	2.11	1.73	1.93	1.69
worst	42.12	74.76	31.02	41.48	56.31
stdev	6.84	9.97	6.78	8.83	10.27

Table 1-4. Forecast accuracy comparison of different forecasting methods measured as Mean Absolute Percentage Error (MAPE) – energy usage

	energy usage				
	STL	Holt-Winters	ARIMA	MLP-ANN fixed lag	MLP-ANN auto
n	63	63	63	63	63
mean	10.14	10.43	10.43	15.66	13.37
best	2.16	2.50	2.50	2.32	2.45
worst	37.62	37.56	37.56	122.94	89.29
stdev	5.90	6.00	6.00	17.35	12.07

The order of ARIMA models (chosen by the software automatically) differ greatly, showing that the seasonality pattern of the performance indicators and individual sites are not uniform. There were 14, 14 and 20 different models used for the three--performance indicators (extract loss, water usage and energy usage respectively) - 63 site combinations. Although the most common model was the ARIMA(0,1,1) - which is equivalent to simple exponential smoothing - only 13/63, 13/63 and 10/63 forecasts were based on this particular model. Majority of the ARIMA models (42/63, 46/63, and 42/63) had a first order differenced component, indicating, that the majority of sites had a trend (improvement) for all three performance indicators. Finally, 21, 30, and 34 models had seasonal components of different kind, confirming that energy usage was more seasonal than water usage and extract loss was being the least seasonal.

Conclusion

The results indicate - answering research question 1 - that both statistical (SLT, Holt-Winters and ARIMA) and MLP-ANN forecasting methods with automatic model parameter selection could be capable of forecasting automatically improvements for several performance indicators and for multiple sites. These algorithms are well advanced to be able to let them run in automatic mode and still achieving acceptable forecasts, even for inhomogeneous sets of time series. This could enable to forecasting multiple sites' performance improvements with limited human intervention. The accuracy of these methods however - answering research question 2 - are not perfect, they are in the region of 10%, measured as Mean Absolute Percentage Error (MAPE). This inaccuracy may be attributable to the fact, that different sites and different performance indicators exhibit very different patterns, that the automated solutions could model only to a certain extent. The results show, that the accuracy of these methods heavily rely on selecting the appropriate models: the right order for the ARIMA model or the right time-lag input vector for the MLP-ANN network. This is particularly visible for the MLP-ANN models: the manually chosen time-lag input vector would always deliver less accurate forecasts, even if higher number of input parameters are used. The limiting factor for these forecasting methods – answering research question 3 – is when past patterns not continuing to the future, or having sudden peaks, high inherent random error, changes in seasonality profile or the lack of a clear trend in the time series. The latter would affect particularly the MLP-ANN models, through the risk of over-fitting, while the former could only be handled through human intervention of reviewing individual forecasts. The study has certain limitations, that only five methods have been evaluated, and model parameters have only been fine-tuned to a limited extent. Further research could evaluate if significantly better forecasting performance than 10% MAPE could be achieved as well as if the ranges and the outliers could be reduced by further tuning the models.

References

[1] Box, G.E.P. and Jenkins, G.M. (1970), Time Series Analysis: Forecasting and Control, Prentice-Hall, Upper Saddle River, NJ.

[2] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015), Time Series Analysis: Forecasting and Control, John Wiley & Sons.

[3] Canova, F. and Hansen, B.E. (1995), "Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability", Journal of Business & Economic Statistics, Taylor & Francis, Vol. 13 No. 3, pp. 237–252.

[4] Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990), "STL: A seasonal-trend decomposition procedure based on loess", Journal of Official Statistics.

[5] Crone, S.F. and Kourentzes, N. (2010), "Feature selection for time series prediction - A combined filter and wrapper approach for neural networks", Neurocomputing, Elsevier, Vol. 73 No. 10–12, pp. 1923–1936.

[6] Fildes, R. (1992), "The evaluation of extrapolative forecasting methods", International Journal of Forecasting, Vol. 8 No. 1, pp. 81–98.

[7] Gardner, D.R. (2017), "STL Algorithm Explained: STL Part II", available at: http://www.gardner.fyi/blog/STL-Part-II/.(accessed 18 March 2018).

[8] Guoqiang Zhang, B., Patuwo, E. and Hu, M.Y. (1998), "Forecasting with artificial neural networks: The state of the art", International Journal of Forecasting, Vol. 14, pp. 35–62.

[9] Holt, C.C. (1957), "Forecasting seasonals and trends by exponentially weighted moving averages", Office of Naval Research Memorandum, Vol. 52, doi:10.1016/j.ijforecast.2003.09.015.

[10] Holt, C.C. (2004), "Forecasting seasonals and trends by exponentially weighted moving averages", International Journal of Forecasting, Elsevier, Vol. 20 No. 1, pp. 5–10.

[11] Hornik, K. (1991), "Approximation Capabilities of Multilayer Feedforward Networks [J]", Neural Networks, Vol. 4 No. 2, pp. 251–257.

[12] Hyndman, R.J.; Koehler, A.B.; Snyder, R.D.; Grose, S. (2002), "A state space framework for automatic forecasting using exponential smoothing methods", International Journal of Forecasting, Vol. 18 No. 3, pp. 439–454.

[13] Hyndman, R.J. and Athanasopoulos, G. (2014), Forecasting: Principles and Practice, OTexts.

[14] Hyndman, R.J., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., et al. (2018), "{forecast}: Forecasting functions for time series and linear models", available at: http://pkg.robjhyndman.com/forecast.

[15] Hyndman, R.J. and Khandakar, Y. (2008), "Automatic time series forecasting: the forecast package for {R}", Journal of Statistical Software, Vol. 26 No. 3, pp. 1–22.

[16] Hyndman, R.J. and Koehler, A.B. (2006), "Another look at measures of forecast accuracy", International Journal of Forecasting, Vol. 22 No. 4, pp. 679–688.

[17] Kourentzes, N. (2017), "{nnfor}: Time Series Forecasting with Neural Networks", available at: https://cran.r-project.org/package=nnfor.

[18] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992), "Testing the null hypothesis of stationarity against the alternative of a unit root", Journal of Econometrics, Vol. 54 No. 1–3, pp. 159–178.

[19] "London Stock Exchange". (2015), http://www.londonstockexchange.com/exchange/news/market-news/market-news-detail/SAB/12576961.html (accessed 11 March 2017).

[20] Massoudi, A. (2016), "AB InBev/SABMiller deal to yield $2bn in fees and taxes", Financial Times, London, available at: https://www.ft.com/content/400e2334-6b6b-11e6-a0b1-d87a9fea034f. (accessed 11 March 2017).

[21] Ord, K., Fildes, R.A. and Kourentzes, N. (2017), Principles of Business Forecasting, 2nd editio., Wessex Press Publishing Co., New York, NY, USA.

[22] R Development Core Team. (2008), "R: A Language and Environment for Statistical Computing", Vienna, Austria, available at: http://www.r-project.org.

[23] Wang, X., Smith, K. and Hyndman, R. (2006), "Characteristic-based clustering for time series data", Data Mining and Knowledge Discovery, Vol. 13 No. 3, pp. 335–364.

[24] Winters, P.R. (1960), "Forecasting Sales by Exponentially Weighted Moving Averages", Management Science, Vol. 6 No. 3, pp. 324–342.

A Beginner's Guide to Open Data: A Case Study

CSABA CSÁKI
Corvinus University of Budapest, Hungary
eMail: Csaki.Csaba@uni-corvinus.hu

ABSTRACT

The latest trend in Open Government Data (OGD) is based on economic interest, namely the idea of innovative, commercial reuse of public sector information. However, reusing open data is not a straightforward exercise. Although there are reports of OGD quality issues, there appears to be no research how newcomers should approach opportunities of OGD reuse. This report highlights basic points related to working with open data for the first time which are based on the author's own experience over cases of typical public sector datasets published with the intent to be reused. The logic of the guideline is built on the main dimensions of assessing OGD quality.

Introduction

The idea of 'open data' has been around for over a decade [23], but the term has special meaning in the public sphere. While open data originally meant scientific or private data, in the context of the public sector the expectation for governments to publish their data is rooted in the principle of 'right to information' [24], [14]. Open government data (OGD) also implies the combined results of various initiatives that are aimed at making public data and information available including transparency and accountability movements [15], e-Government [4], open government [17], or data reuse [8].

The reuse of data originally generated in or utilized by public sphere entities brought to light new issues and requirements related to the quality of such data as published. Quality in general, or data quality in particular are peculiar, hard to define concepts, and this is not easier in the context of OGD either. While there are numerous data quality frameworks [2] and open data quality assessment approaches [11], they tend to focus on a set of specific dimensions along which quality may be measured and might even provide guidelines how to improve quality or avid problems [26], but their interest does not cover offering help to newcomers of open data. Consequently, first time open data users are often unprepared for what awaits them and how to prepare for a successful OGD reuse project.

Therefore, the aim of the research reported here was to understand the typical data quality challenges users of OGD may expect to face and to provide some ideas how to prepare in order to resolve them. To establish a context, this paper first reviews the key ideas behind open government data and its secondary utilization, then offers a summary of the literature on (open) data quality. The methodology section sets the research questions and the path selected to answer them. This is followed by an analysis of the data reuse process based on dimensions of quality. The main body of the paper describes cases of data problems data are typical during the above typified reuse process. The paper concludes with a summary of theoretical and practical significance.

Open Data of the Government

Open data is usually defined as data that is made available to be freely accessible and reusable [23]. While in technical terms data differs from information – the former being a term related to the storage and preservation of symbols (in itself having no meaning), while the latter referring to data interpreted by an actor in a given context [20] –, reusing open data typically means contextual matching, which is thus interpreted as information by the end user. From this point of view there appears to be

little differentiation between data and information in the OGD literature. Therefore, public sector information (PSI) is defined as "*information, including information products and services, generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for a government or public institution*" ([21], p. 4.) (see also [14]). Consequently, availability and access to such information in various forms falls under both 'open data' initiatives [1] and has fallen more broadly under the umbrella of Electronic Government [13]. Most importantly, from the point of view of information technology, more and more data is being made available in bulk over the Internet for others to use and repurpose [31]. With respect to such data, especially for data originated in the public sector or published by public entities for that purpose, there are some arguments for them to be published to serve economic interest [16],[34]. However, with the push for commercial reuse [34] – including integration with other, existing datasets (e.g. to form mashups, [2]) – the focus becomes how *value added* services may be created [33].

The basis for value added services built on open data is the assumption that users of open data seek information which may or may not be readily available in the published data set [5]. Providing analytical capabilities is one potential value generator [28]. Another ethos attached to open data is its ability to drive innovation [16]. One important option in such innovation is to connect various datasets thereby creating mashup-like visualizations or generating new insights [18]. The idea of innovative, commercial reuse of public sector information underlines the question of what role quality plays in such processes [8].

Quality Frameworks

With the rapid evolution of the Internet the number of informational outlets has grown quickly especially over the last few years. As a result, low information quality (IQ) is one of the most challenging problems for consumers of data. However, 'quality' is difficult to define especially in the context of such a 'soft' product as data [32]. Furthermore, scholarly articles on IQ define their key concept in varying ways [3].

Over time, started as early as the 80s, numerous data and information quality frameworks have been proposed, mostly based on the idea of 'assessment' of quality using a set of predefined dimensions [6]. The *technical view* associates information quality with the accuracy of the data stored in databases [19] and looks at issues surrounding timeliness of update, system reliability, system accessibility, system usability and system security. Another *technology focused approach* considers machine readability (such as linking, finding, relating and reading data using automated processes), and characteristics include number of formats, traceability, automated tracking, use of standards, trustworthiness, authenticity or provenance [27]. Perhaps the most commonly used simple definition of *user side IQ* interprets the term as "fit-for-use" [32]. However, IQ defined this way remains a relative construct whereby data considered appropriate for a given use may not display acceptable attributes in another setting [29]. Furthermore, fit-for-use does not immediately allow for ready measurability and it requires additional detail in order to be operationalized [11]. However, quality of the data as stored, accessed and manipulated can substantially differ from the quality of the information that the data contains or that the data can offer in terms of information gleaned from it. This might be addressed by adopting a measure of *perceived information quality* (PIQ) that could be based on the assessment of the ability of intended audiences to generate new information and gain additional insights from the data. In the user centric perspective of the Internet this would translate into quality stemming from the degree to which information is suitable for doing a specific task by a specific user in a certain context [10].

Defining and assessing Open Government Data Quality (OGDQ) may require a slightly different approach that may be distinguished from general data quality discussions. For example, open government data must have some level of trust associated with it that is derived from authority, and trust in open data may anchor its value proposition in accountability [22]. One additional concern is the availability and accessibility of various types of data combined with the question of timeliness (i.e.

whether data are out-of-date): for example, the Open Data Barometer [7] monitors the availability of open public datasets around the world using specific measures. On the other hand, focusing on technical aspects does not require special characteristics in case of open data. However, while the *Linked Open Data* (LOD) approach concentrates on provenance that may enrich the context of open data [26], quality assessment dimensions applied under LOD not only include origin, attribution, traceability, accessibility and presentation but reliability and trustworthiness as well.

In conclusion, while there are many quality frameworks and specific cases of individual issues, there appears to be no systematic overview of all potential issues and how they may impact the research of newcomers to open data – and what to do about it. No measures, characteristics, or dimensions may guide a new practitioner who bravely ventures into the realm of utilizing open government data, the subject to which this paper now turns.

Research Question

The intention is to adopt and apply ODQ theoretical frameworks to a specific important open dataset following the process of exploring new data, therefore the Research Question proposed is: 'How the various frameworks to assess the quality of open data may be used to prepare beginning researchers for the task of empirically approaching an actual open data dataset?'. The question is addressed in two steps: first the assessment frameworks and their proposed dimensions are used to establish an open data discovery process, then these steps are used on a representative case study to demonstrate how beginners may attack the challenge of open data. The process is recreated from the literature discussed so far in the (first half of the) paper, while the case used is the Tender Electronic Daily (TED) csv open dataset containing European Union (EU) public procurement (PP) data.

The TED dataset is comprised of public procurement data published daily as the official online version of *Supplement 32 to the Official Journal of the European Union* and individually accessible as part of the TED public procurement website (http://ted.europa.eu/). The stated goal is to make European public sector procurement more accessible and PP calls made in EU member states whose value falls above minimum threshold amounts (stipulated in EU regulation) ought to be published in the TED. Five affiliated countries also publish tender and award notices in the TED Journal to gain access to the EU market. Data in the Journal are collected from standardized public procurement forms as required by the corresponding EU Directives (2014/17 and 18) and their Annexes. As part of the open data initiative of the EU this data is republished annually containing both tender calls (notices) and result announcements (awards)[1]. Each annual dataset is published in csv format using UTF-8 coding.

The illustrative example of this research uses the TED csv data. This open data set is very complex as there are three levels of procurement information: a) contract notices (CN); b) contract award notices (CAN); and c) contract awards (CA). While the process of public procurement is inherently complicated, for now it should suffice to state that one (or occasionally two) CNs lead to one CAN, but one CAN may lead to one or more CAs associated with it (this is because a CN may have a preliminary notice, while a single call may have several parts or lots with each leading to a separate contract being awarded under the same CAN). All data files were downloaded January 17, 2017 and analyzed using MS Excel and Access (2010), SPSS (v22.2), and Oracle DB (11g Release 11.2.0.4). The datasets are accompanied by a codebook [30] that serves as a guide: CN datasets have 54 fields, while CAN/CAs have 50 fields (some fields are not from the original forms but generated during csv output). Due to documented concerns over maintaining consistency of reporting and the procurement standard forms from which the data are captured, only data from 2009 to 2015 is considered in this study. The total size of the fourteen different data files was approximately 2.13 GB consisting of over 4.5 million records.

[1] (https://data.europa.eu/euodp/en/data/dataset/ted-csv)

Open Data Quality Dimensions

Based on the key dimensions of the open data quality frameworks discussed above, an ideal process was constructed that follows the main steps of discovering and using open (government) data. Each step has its quality challenges for new users to be discussed in the next section.

- **Awareness and availability**: Quality of OGD is increasingly measured by its availability [7]. Indeed, more and more public sector data has been made available and linking datasets is becoming the norm making it easier to find relevant datasets.
- **Accessibility**: Data being published does not automatically mean it can actually be accessed directly or for free [11]. One of the main arguments for OGD is the possibility of it being repurposed in value added services, which assumes unhindered access and an open license [1].
- **Readability**: Once data has been located, an important quality measure of OGD is whether it is in machine readable format [31]. Without it users cannot take advantage of modern technologies that can locate, read and process data automatically. Indeed, connecting structured and machine readable data can be semantically queried allowing for advanced Big Data analytics [9].
- **Technical qualities of the data**: This is to be assessed as the next step before data may be utilized. Raw data has no value in itself, value is generated by technical and economic capabilities over time [12]. It is strongly influenced by the existing technology infrastructure of both producers and users, but high technical quality is an essential ability to design a service of high end user quality.
- **Content and structure**: Before use, it is necessary to assess the quality of the data in relation to its content and domain context. Linking data sets as part of commercial service application requires the matching of data based on their meaning within the application domain. This forms the bases for multivariate analysis that may help in obtaining new insights [5].
- **Traceability**: On top of linking datasets, it is very important to be able to check the validity of data before use. There are dedicated standards and best practices how to fulfil provenance requirements of OGD quality on behalf of the issuer [26],[27].
- **Usability**: Once data is reviewed, the next step is its actual use in presentations, analysis or service creation. Ease of use depends on the (language and date) formats, the tools required to work on the data, and the structure of the data in the files. In addition, subject knowledge and deep understanding of the domain may be required to effectively and efficiently utilize the data [25].
- **Fit-for-purpose**: Even with a high quality 'mechanical' usability, the final question is whether there is value to be generated from it. The ability of end users to generate value is influenced by the complexity of the dataset, its coverage, its content (as earlier) and its representation. As summarized by [8] data should not only be "*intrinsically good, but also contextually appropriate for the task, [and] clearly represented*", which also implies that "*information may be … completely inappropriate for [certain users who] have different temporal, security, granularity, or other requirements*" (p. 57).
- **Feedback**: Although this is not part of the quality dimensions, several authors (e.g. [2],[28],[31]) emphasize the importance for users of open data to provide feedback to owners of datasets in order to allow for improvements along various dimensions (what to release and how).

Utilizing Open Data

Awareness and availability

The case: The TED dataset is being issued in several formats: individual pdf, daily xml, monthly xml and annual csv – so one has to look out for the latest release and choose the time-span needed. In addition, every few years the data structure is updated. Indeed, TED data fields have been majorly reorganized in 2017 (after this research had started) making it necessary to not only download the up-

dated versions, but also to modify algorithms and code pieces used to analyze the data.

The note for beginners: Due to the large number of available datasets it is now more difficult to find relevant data in the increased volume of output. Also, one has to constantly be on alert regarding the freshness of the data – partly because of outdated links, partly due to sources being updated regularly. It is recommended to check for the data source to be authentic and whether the data came with description/documentation (which should also be up to date).

Accessibility

The case: The EU PP data in the TED is made available for free and is accessible in various standard formats. The granularity of the TED csv data at file level is one year, but integrated files covering several years are also available. Each record is one notice or one award. Individual file sizes span from 100kB to 1.6GB. None of these posed any issues during download. The official TED Journal, however, also offers individual notices as well as daily and monthly digests (in zip-ped xml format) – the size of which is typically 6-7 MBs per day and 150-200MBs per month.

The note for beginners: One should start with checking for potential fees and licensing requirements. Download speed should generally not be a problem, except for very large (Big Data) files. Therefore, the most important factor during download is granularity. In some cases the units released as OD are at a low level and thousands (or even tens of thousands) of small records need to be accessed and pieced together – at the other end of the spectrum data may be released at a high level of aggregation resulting in massive file sizes and complex datasets.

Readability

The case: Opening the TED csv files in Excel resulted in a few surprises such as scrambled lines – due to language settings of the OS, which could be rectified by resetting the default separator. However, there were additional problems with the UTF-8 encoding as well as it is language dependent – and each tool used had its own way of dealing with this problem. Since EU members may use any of the official languages for their PP tender announcements, basic UTF-8 reading (such as English as default) results in scrambled characters for languages like Greek, Hungarian, Swedish, etc. It turned out, that the UTF-8 setting in the analysis/DB tools should allow for 'all' languages for text from every EU language to be read and displayed properly (while MS Excel required the "import" function for proper reading of UTF-8).

The note for beginners: Even though data are now usually published in non-proprietary formats using standard encoding schemas, language differences could cause interpretation problems depending on the tool used. CSV does not carry language information, but UTF-8 requires a so called BOM character for font mixing (i.e. for characters from various language sets to be displayed properly). One should make sure that several tools are available and that their settings fit the requirements of the data format – if something does not look right, it is wise to try different language, coding and location settings or, alternatively, change tool.

Technical qualities of the data

The case: Checking various formatting aspects of the data resulted in several necessary transformations. It might sound banal, but date formatting is not straightforward in most tools and they might not recognize all date formats – and extra coding was required to transform all dates into acceptable formats (for example from "DD.MMM.YY" or "DD-MMM-YY" to "YYYY.MM.DD" depending again on language. Another technical issue concerned the length of text fields as different tools have different default limits. For example, while Excel had no problem with lengthy text fields, Access would truncate fields with longer size, while Oracle would reject such records (and choosing a very large value for such fields would result in a much larger database file requiring more storage). One of the main issues here is that csv datasets do not carry datatype information.

The note for beginners: Most public sector data is stored in databases, but during publication they are stripped of any datatype information, yet (different) types might be automatically assigned during utilization depending on the tools used. Although Excel has a limited capability to differentiate between a few datatypes such as Text, Date or

Number, it would assign these automatically and often reverts to the "General" type as a default. Other tools, such as databases or data management applications would offer a range of types which might be quite sophisticated but would require manual assignment (of both types and field sizes – and knowing the longest possible text field is important). So again, try different tools and settings to see which helps.

Content and structure

The case: Checking the content turned out to be the most complicated part of the discovery process. All in all, there were missing field values, unexplained duplication of records, and records that did not represent actual PP processes (as turned out to be through manual analysis of the original TED data). In addition, there were fields with multiple values, indicating that there was no semantic checking during the filling out of the notice forms. Furthermore, certain types of procedures, such as "periodic indicative notice without a call for tender" had been duplicated in the csv files – some of them being exact copies, others having small differences in one field or another. Finally, 'cancellation' of notices are not always indicated consistently: some of them resulted in a new record, others had a modified value in the corresponding field. All these impacted the analysis and required some decisions what to include in the statistics and what to leave out.

The note for beginners: Public sector data is typically captured through forms and templates, which would enforce certain restrictions, but relying on this could lead to problems as any programming or deliberate mistakes when filling out the forms could result in data quality issues. Such errors may only be discovered using deep domain knowledge and understanding of the context. It is necessary to consider missing data or duplicates. It is not unusual to see multiple values entered into one field. Use the guideline if available, but still be careful and be wary of missing data or errors.

Traceability

The case: The csv files downloaded contained no direct link to the actual TED notice a given record is generated from. Instead, checking validity of samples required manual look-up in the TED search tool. Furthermore, each call notice needed to have either a cancellation or one or more corresponding awards – but this backward link had been replaced by projected future link (instead of the CAN pointing which CN it is resulted from, the CN offered 'future' CAN values).

The note for beginners: Open data is rarely standalone and is often composed of several parts or related datasets: it has (should have) references within the area covered, but it might use references to other sets (e.g. country codes, national abbreviations, etc.). Despite the available standards, the actual reliability of data set linkages varies widely hindering the users' ability to connect relevant components of related datasets. So pay special attention and double check all such references for correctness.

Usability

The case: Although the structure of the TED csv files is described in a guide, understanding the meaning of various fields required in depth understanding not only of public procurement but also the specific details of EU procedures. This was further complicated with the fact that the csv fields did not fully reflect either the fields in the TED DB or the original forms contracting authorities use when submitting data related to calls and results. Even the guide did not explain the mapping between these three formats – requiring new effort of connecting the dots when a new research question was asked from the datasets.

The note for beginners: Ease of use depends on three things (each of which has already appeared above): the format of the file (csv); the tools required to work on the data (various tools should be tried); and the structure of the data in the files. To turn the raw open data into a useful asset, it is typically necessary to execute manual cleansing and restructuring to make it digestible by analytical tools. So check formats, carefully select tools to use (they do differ in their ability to manage various formats), and if necessary, apply structural transformations (ones, that do not alter content, only structure).

Fit for purpose

The case: As the original research was aiming at statistical analysis of EU PP data, a very intimate knowledge of European public procurement for-

malities appeared essential. One peculiarity was related to the non-standardized way how individual countries have inputted data into the TED DB: some were at contracting authority level while other countries exercised control at the central government level, resulting in missing codes, or missing values or inconsistencies in the names of authorities – all impacting statistical analysis over or involving affected fields. There was one additional recurring issue that affected the analysis of the datasets, the problem of missing values. In some cases a simple visual browsing of the file was enough to see that certain records had missing values, while in other cases the statistics revealed a number of "no value" items.

The note for beginners: The ultimate goal of OGD reuse is value generation. Squeezing value out of public data often requires the ability to deal with the complexity of data sets – especially with merged or linked datasets. Preparing datasets for statistical analysis demands time and expertise as filtering out problematic fields and records requires deep domain knowledge. It might even be necessary to reconsider the questions that could be meaningfully answered from a given dataset. Any decisions made about how to transform certain pieces of data obviously impacts on statistical analysis results.

Feedback

The case: During the exploration of the case data a member of the PP data analysis project contacted the issuer of the data (TED) partly to ask for clarification of uncertainties partly to inform them about some findings related to the quality of the csv dataset. There were a few emails exchanged resulting in not only a better understanding of the dataset structure but also leading to changes of the data-structure and new documentation on the issuer side.

The note for beginners: It is strongly recommended to engage with the owner (or issuer, manager) of the data being utilized. This could lead to mutual benefits through clarification of issues and potential improvements.

Conclusions and future research

The paper presented research results of a case study on the reuse of open government data with the dedicated goal to provide help to newcomers to this area, who intend to utilize OGD in providing value added services. The main message is that the nature of public sector data and the mechanisms used during their production and publication are specific to the context and do impact on the quality of the data being disseminated. This has to be considered when utilizing PSI or OGD for innovative services. Potential future research involves the categorization of root causes that hamper OGD quality. This further requires an understanding of the differences between the OGD and the private data ecosystem, which is also planned to be investigated.

Acknowledgement

This work was created in commission of the National University of Public Service under the priority project KÖFOP-2.1.2-VEKOP-15-2016-00001 titled „Public Service Development Establishing Good Governance" and of the Corvinus University of Budapest. The author wishes to thank the staff of the European Commission for the use of the TED csv dataset and also acknowledges the contributions of Prof. Eric Prier and Prof. Clifford McCue of Florida Atlantic University.

References

[1] Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L. and Wouters, P. (2004): An international framework to promote access to data, *Science*, 303(5665), pp. 1777-1778.

[2] Attard, J., Orlandi, F., Scerri, S. and Auer, S. (2015): A systematic review of open government data initiatives, *Government Information Quarterly*, 32(4), pp. 399-418.

[3] Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009): Methodologies for Data Quality Assessment and Improvement, *ACM Computing Surveys*, 41(3), pp. 16-52.

[4] Bertot, J.C., Gorham, U., Jaeger, P.T., Sarin, L.C. and Choi, H. (2014): Big data, open government and e-government: Issues, policies and recom-

mendations, *Information Polity,* 19(1, 2), pp. 5-16.
[5] Cavanillas, J.M., Curry, E. and Wahlster, W. (Eds.). (2016): *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, Cham: Springer International.
[6] Chai, K., Potdar, V. and Dillon, T. (2009): Content quality assessment related frameworks for social media. In *Proc.s of the Computational Sci. and its Applications Conf.,* Springer, Berlin, pp. 791-805.
[7] Davies, T. (2013): *Open data barometer*: 2013 global report, World Wide Web Foundation and Open Data Institute.
[8] Dawes, S.S. and Helbig, N. (2010): Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency, in: Wimmer, Chappelet, Janssen, and Scholl (Eds) *9th IFIP 8.5 Conference on Electronic Government*, Springer LNCS-6228, pp. 50-60.
[9] Dwivedi, Y.K., Weerakkody, V., Janssen, M., Millard, J., Hidders, J., Snijders, D., Rana, N.P. and Slade, E.L. (2015): Driving innovation using big open linked data (BOLD) – panel. *In Lecture Notes in Computer Science*, Vol. 9373, pp. 3-9.
[10] Emamjome, F.F., Rabaa'i, A.A., Gable, G.G. and Bandara, W. (2013): Information quality in social media: a conceptual model. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2013), Seou*l, AIS Electronic Library (AISel).
[11] Frank, M. and Walker, J. (2016): User centred methods for measuring the quality of open data. The *Journal of Community Informatics*, 12(2), pp. 47-68.
[12] Hjalmarsson, A., Juell-Skielse, G. Ayele, W.Y., Rudmark, D. and Johannesson, P. (2015): From Contest to Market Entry: A Longitudinal Survey of Innovation Barriers Constraining Open Data Service Development, *ECIS 2015*, Paper 78. http://aisel.aisnet.org/ecis2015_cr/78.
[13] Jaeger, P.T. (2003): The endless wire: E-government as global phenomenon. *Government Information Quarterly*, 20, pp. 323-331.
[14] Janssen, K. (2011): The Role of Public Sector Information in the European Market for Online Content: A Never-Ending Story or a New Beginning? Inf*o: the J. of Policy, Regulation and Strategy for Telecommunications, Inf. and Media*, 13(6), pp. 20-29.
[15] Janssen, K. (2012): Open Government Data and the Right to Information: Opportunities and Obstacles, *The Journal of Community Informatics*, 8(2).
[16] Jetzek, T., Avital, M. and Bjorn-Andersen. N. (2014): Data-driven innovation through open government data, *Journal of theoretical and applied electronic commerce research*, 9(2), pp. 100-120.
[17] Kassen, M. (2013): Globalization of E-government: Open Government as a Global Agenda, Benefits, Limitations and Ways Forward, *Information Development*, 30(1), pp. 51-58.
[18] Lebo, T., Erickson, J.S., Ding, L., Graves, A., Williams, G.T., DiFranzo, D. and Shangguan, Z. (2011): Producing and using linked open government data in the twc logd portal, in: *Linking Government Data*, pp. 51-72.
[19] Levitin, A. and Redman, T. (1995): Quality dimensions of a conceptual view. *Information Processing & Management*, 31(1), pp. 81-88.
[20] Lindman J., Rossi, M. and Tuunainen, V. (2013): Open Data Services: Research Agenda, in: *Proceedings of HICSS-46*, pp. 1239-1246.
[21] OECD [Organisation for Economic Co-operation and Development] (2008): *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* [C(2008)36]. Online, Downloaded December 10, 2016: http://www.oecd.org/internet/ieconomy/40826024.pdf.
[22] OECD [Organisation for Economic Co-operation and Development] (2017): Trust and Public Policy: How Better Governance Can Help Rebuild Public Trust, *OECD Public Governance Reviews*.
[23] Open Knowledge Foundation (2006): *Open Knowledge Definition*. Online, letöltve 2017. szeptember 7: http://www.opendefinition.org/.
[24] Parks, W. (1957): The open government principle: applying the right to know under the constitution, *The George Washington Law Review*, 26(1), pp. 1-22.
[25] Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S. and Carlson, D. (2011): A conceptual framework for managing very diverse data for complex, interdisciplinary science, *Journal of Information Science*, 37(6), pp. 555-569.
[26] Pignotti, E., Corsar, D. and Edwards, P. (2011): Provenance Principles for Open Data. in: *Proceedings of DE2011*.
[27] Rula, A. and Zaveri, A. (2014): Methodology for assessment of linked data quality, in: *Proceedings of the 1st Workshop on Linked Data Quality at the 10th Int. Conference on Semantic Systems*.

[28] Susha, I., Janssen, M. and Verhulst, S. (2017): Data collaboratives as "bazaars"?: A review of coordination problems and mechanisms to match demand for data with *supply Transforming Government: People Process Policy*, 11(1), pp. 157-172).
[29] Tayi, G.K. and Ballou, D.P. (1998): Examining data quality. *Comm.s of the ACM*, 41(2), pp. 54-57.
[30] TED. 2016. *TED Processed Database: Notes & Codebook*, Version 2.2. Online, Downloaded Jan. 20.01.2017: http://data.europa.eu/euodp/repository/ec/dg-grow/mapps/TED(csv)_data_information.doc.
[31] Ubaldi, B. (2013): Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives, *OECD Working Papers on Public Governance*, No. 22. Online, Downloaded 20,01. 2017: http://dx.doi.org/10.1787/5k46bj4f03s7-en.
[32] Wang, R.Y. and Strong, D.M. (1996): Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), pp. 5-33.
[33] Zeleti, F.A., Ojo, A. and Curry, E. (2016): Exploring the economic value of open government data." *Government Information Quarterly,* 33(3), pp. 535-551.
[34] Zuiderwijk, A., Helbig, B., Gil-García, J.R. and Janssen, M. (2014): Special Issue on Innovation through Open Data – A Review of the State-of-the-Art and an Emerging Research *Agenda, Journal of Theoretical and Applied Electronic Commerce*, 9(2), pp. 1-8.

Big Data Analytics Possibilities for the Space Domain

[1]László Bacsárdi – [2]Gergely Bencsik – [3]Zoltán Pődör
Institute of Informatics and Economics, University of Sopron
eMails: [1]bacsardi@inf.uni-sopron.hu, [2]bencsik@inf.uni-sopron.hu, [3]podor@inf.uni-sopron.hu

Abstract

In the 21st century, our everydays' operations are centralized around the information. Different Big Data analyzing methods are applied in various domain to provide enormous capacity of data. In the last years, space became a relevant source of Big Data. Not only the space telescopes and probes are producing data, but high amount of data is transmitted from the satellites orbiting Earth as well. Due the four V's of Big Data, the volume, high velocity, the data variety as well as the veracity of these data, the space industry is facing several challenges. In this article, we present some of the state-of-the-art solutions and detail our self-developed methodologies which can be applied in the space domain.

Introduction

Since the beginning of the space age in 1957 with the launch of Sputnik 1, the world's first satellite, space plays a crucial role in our activities. The Global Navigation Satellite Systems (GNSS) provide not only accurate positions while driving our car but foster innovation in various domains including truck movement in the precision agriculture and different time synchronization services. Although we refer those systems as GPS, there are various GNSS systems like the American GPS, the Russian Glonass, the Chinese Beidou and the European Galileo system. Another evolving field is the Earth Observation (EO), where different Earth orbiting satellites provide high definition images of our planet. In this article, we would like to focus on European activities since Hungary became full member of the European Space Agency (ESA) in 2015. In the framework of the Copernicus Earth Observation program, ESA recently launched the Sentinel satellites. So far four Sentinel satellites have been placed in orbit (Sentinel-1A and 1B, Setinel-2A, Sentinel-3A) and further satellites are planned to be launched in the coming years in 9 different sentinel missions. Sentinel-1 satellites

provide day and night radar imaging for land and ocean services. Sentinel-2 provides high resolution optical images for land services and for emergency services, while Sentinel-3 provides ocean and global land monitoring services. These satellites are providing terabytes of data every day [1]. Methods of Big Data analysis will play an important role in the coming years since data acquired from space fulfils the four V's of the Big Data concept [2]:

Volume: the main characteristic that makes data big is the huge volume. Nowadays, when the biggest part of the data is generated by machines, networks, sensors and human interactions (e.g., social networks), the volume of the data grows exponentially every year.

Velocity: it deals with the frequency of the incoming data. The real time data flows (e.g., data from sensors) are usually massive and continues. The velocity is connected to two problems: the rapidly increasing speed of the incoming data, and the online/offline analyzation problems of this huge datasets.

Variety: with increasing volume and variety comes increasing variety. Datasets collected in the space domain come from a diverse set of sources, they may be both structured and unstructured, but even the structured elements have a wide variety regarding their structural features. Traditional, structural data types are usually stored in relational databases. Unstructured data (e.g., images, twitter feeds, audios, web pages, etc.) is a fundamental concept of Big Data. One of the biggest aim of Big Data is to use the adequate technology to take this unstructured data and make sense of it.

Veracity: it refers to the biases, noise and abnormality in data. Before the using of data, they must be clean, consist and consolidate.

In the past years, we developed two techniques which can be applied in Big Data analysis. The CReMIT (Cyclic Reverse Moving Interval Technique) method [3] extends the analysis possibilities of periodical time series to find more precise correlations by creating derived, secondary time series as dependent or independent variables. Using given CReMIT parameters, a huge number of the derived time series can be created which are not depend on the further used analyzing methods. But based on our experiments, the linear and non-linear analyzing possibilities can cause such big number of analysis possibilities with which the results are born just randomly, despite of the exact mathematical environment, and the results are misidentified as real correlations. The Random Correlation method helps us to analyze and determine the random level of given results [4, 5].

Market of the Big Data from space

The Earth Observation (EO) market will grow in the next years despite the high initial cost of investment and the strong government control. From 2007 to 2016 more than 180 EO satellites were launched, and over the next 10 years about 600 EO satellites are planned to be launched to support creating high resolution images of our planet. The EO data are efficiently useable in several domains including infrastructure, agriculture, industry, energy and power, navigation.

The EO data and services market should reach $8.5 billion by 2026 based on the current growth prognosis. According to a special value-added services (VAS) model it has a combined market potential of about $15 billion. The largest market of VAS are the infrastructure and national resources monitoring, but the connected solutions are often lower-cost or free data solutions. More than 45 countries are expected to launch satellite capacity, and over 20 of them should have partnership with the private sector; this is expected to generate over $33 billion in manufacturing market revenues.

According to a market study of the Northern Sky Research, the price of the data from space will constantly decrease in the next years both the Synthetic-aperture radar (SAR) segment and the optical segment [6]. (SAR stands for a common approach to create non-optical records.) This decreasing subserves the wide-ranging use of the space data.

There are many companies which are building on new methodologies and different algorithms to handle, analyze and visualize the EO data, to detect patterns and to build predictive analytics. The traditional methodologies cannot cope with this mass

and type of data. Handling the higher frequency collected data, a Big Data environment needs special approaches to open new service areas. It is important to develop special infrastructure and software solutions, which can handle this Big Data based challenge.

The application possibilities of the collected EO data is practically unlimited. There are many community and institutional platforms for this, and they usually are attached to different tasks and problems as it is listed below.

Balhar et al [7] investigated the urbanization by using their special platform, urban Thematic Exploitation Platform which helps to differentiate between urban and rural settlement forms. It is possible to document the changing of built environment based on satellite images. Hame et al. aim was to compute vegetation indices and mapping the land and forest cover to estimate the volume and the change of biomass [8]. Forestry Thematic Exploitation Platform offers a one-stop-shop for forestry remote services.

The devised platforms and technologies can be applied not only on space data, but in other research areas as well. Albani et al. developed a platform providing the possibility to access, process, examine and visualize satellite and collateral datasets [9]. The main services of this platform are: (1) search and download Sentinel images, (2) compare the selected images with change detection, (3) continuous monitoring of the images and (4) search on social networks. The researchers demonstrated the advantages of this platform by automatized data access, process, analysis and visualization. They plan to add further data and services to the platform increasing the number of functionalities made available to the users.

Soille et. al developed a four-layer based platform named JRC Earth Observation Data and Processing Platform (JEODPP) [10]. It is a versatile, petabyte-scale platform to serve the data pretence of different projects and applications. This platform enables the data and information extraction from the stored image datasets using a cluster environment for batch processing. JEODPP provides an opportunity to achieve a web-based remote desktop access with huge number of special software components, and a web-based interactive data analysis and visualization system (Jupyter). First and last, JEODPP is a domain independent platform, which can serve different applications with EO imagine data.

Two new methodologies

The researches based on EO data focuses mainly on the primary data. It means that after storing and cleaning EO data, data is used directly to achieve goals, no derivation or numerical analyses are performed. The new CReMIT (Cyclic Reverse Moving Intervals Techniques) method extends the analyzing possibilities with creating new datarows by different transformations. However, the increased number of datarows eventuate that circumstances, near which correlations can be created randomly. Therefore, Random Correlation theory was developed to make distinction between real and random models.

CReMIT

Searching for relationships between time series is a big challenge of statistics and data mining. There are many possibilities to examine the connection between time series including the correlation and regression analyses methods. The completeness of the examinations does not depend only on the applied analysis methods, but it depends on the sphere of the involved variables (dependent and not dependent) as well. If we have a proper length time series, the temporal changes can be examined by using moving intervals and evolution techniques [11]. The essence of the moving interval technique is that the length of the actual examined time interval is always fixed, and the starting point of them is moved forward. In case of the evolutionary technique, the starting point is fixed and the length of the interval is increased in each step, as it is illustrated in Figure 1.

Based on these two, above mentioned windows-based techniques, a special windows-based method was developed. CReMIT combines the advantages of these two methodologies [3] and makes it possible to systematically widen the

sphere of the used dependent or independent variables by creating new, derived time series in a systematic way. The CReMIT method generates a lot of new, aggregated time series based on the periodicity of the original time series. It is important that the CReMIT is a part of the full-time series analysis process, and it is independent of the further applied analysis methods, as it is shown in Figure 2.

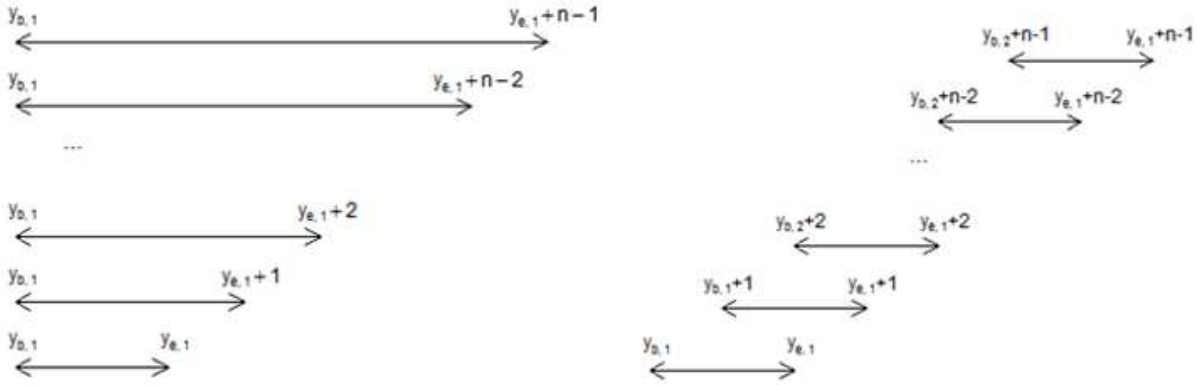


Figure 1: Evolutionary (left) and moving interval (right) techniques

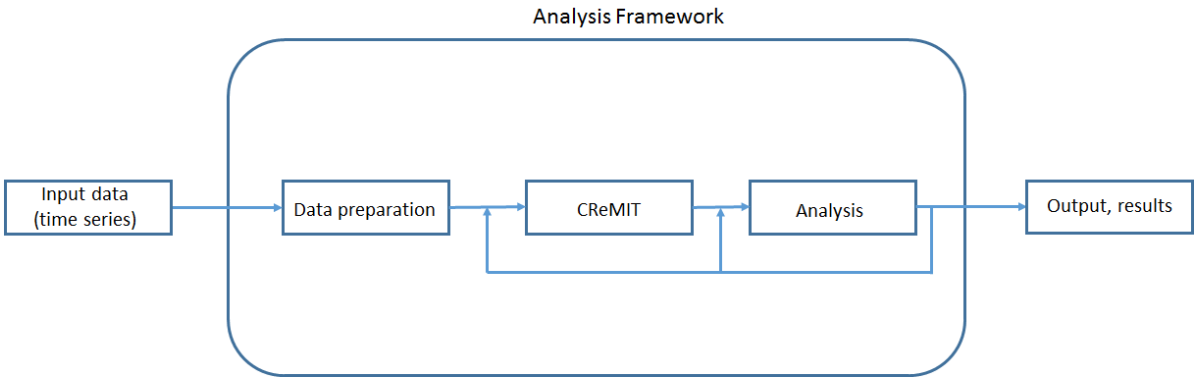


Figure 2: Time series analysis process

Let be given a periodical ts time series with length *m*, and its natural period is denoted by P (for example if ts contains monthly data, than $P = 12$). The elements of ts are stored in a vector, tsv, and the first element of this vector is tsv_1, it is the chronologically latest element of tsv.

$$tsv = \begin{pmatrix} tsv_1 \\ tsv_2 \\ \vdots \\ tsv_m \end{pmatrix}$$

Let us denote by *SP*, where $1 \leq SP \leq P$ the starting point of the currently applied examination, *SP* is always in the first period of *tsv*. This element is the SP^{th} element of *tsv*, based on our notation. There are two special parameters: i and j to define a window above $the\ time\ series tsv$. Parameter i denotes the time shifting, and j the window width values on the basis of tsv indexes. The minimal value of time shifting can be $i = 0$ (it means that first element of the derived time series will be SP, in other words the first element of the actual windows is tsv_P), the maximum value of it are denoted by I.

The window width can be 1, but then, to the further simple using in our notation $j = 0$. The maximum value of j is denoted by J. The parameters i and j depend on the actual examination, and they must be defined by the user before the running of CRe-MIT. A derived window has element in each periods of the original time series. Based on periodicity P of ts the above defined window will be periodically repeated MCN times, where MCN is the maximum cycle number:

$$MCN = \left[\frac{m - (SP + i + j)}{P}\right] + 1$$

where [ ] is the entire function. There are created MCN pieces new windows, based on user defined parameters SP, i and j. The starting and ending points of these windows can be defined as $[SP + i + k * P; SP + i + j + k * P]$, where $0 \leq k \leq MCN - 1$. To store these values, two temporal vectors are defined:

$$index_{begin} = \begin{pmatrix} SP + i + 0 * P \\ SP + i + 1 * P \\ \vdots \\ SP + i + (MCN - 1) * P \end{pmatrix}$$

$$index_{end} = \begin{pmatrix} SP + i + j + 0 * P \\ SP + i + j + 1 * P \\ \vdots \\ SP + i + j + (MCN - 1) * P \end{pmatrix}$$

Based on these vectors and a user defined TR transformation function (e.g., average, summary, minimum, maximum, etc.), which depends on the actual examination, we can create the derived time series, which elements are a special aggregation of the element in the windows.

$$tr\_ts_{SP,i,j} = \begin{pmatrix} TR(index_{begin}[1]; index_{end}[1]) \\ TR(index_{begin}[2]; index_{end}[2]) \\ \dots \\ TR(index_{begin}[MCN]; index_{end}[MCN]) \end{pmatrix}$$

The whole number of transformed, derived time series is defined by the given parameters of ts time series: the length of the original time series, m and the natural periodicity of it, P and by the values of user defined parameters SP, I and J.

Random Correlations

Seeking correlations between datarows is always a difficult task. However, if a correlation is found, the reliability of the correlation must be checked. There are many models and methods to check the endurance of the given result such as R^2 and statistical tests. These methods come from the field of traditional mathematics mainly. However, the theorem of Random Correlations (RC) has another point of view [4, 5]. RC assumes that the increased number of datarows and methods of analysis eventuates nearly countless analysing possibilities and this huge possibility space creates that environment, in which correlations will surly evolve. However, these correlations do not truly exist, they are just born randomly because of the huge number of analysing possibilities. Even if scientists use tests to check the result's endurance, even then they will surely find a good correlations despite of the followed precise research methodologies. From this practical point of view, the possible random property of results can be easily hidden from the scientists. The concept of Random Correlations is summarized in Figure 3..

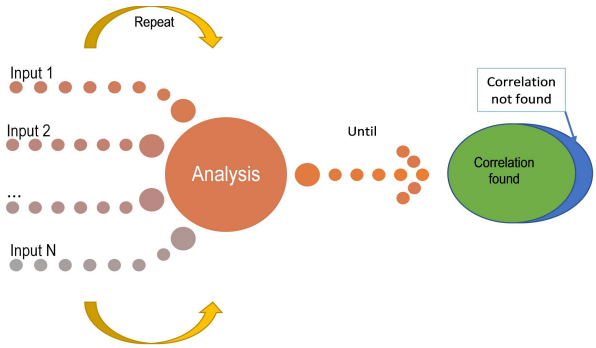


Figure 3: The concept of Random Correlations

As we can see in Figure 3, RC means that repeatedly using more and more inputs and methods of analysis, the probability of "correlated" notation will be very increased. Figure 3 shows that in the set of results, the "correlation found" is highly possible, independent of which "endurance measurement" method is performed. The question is, how such result sets can be produced. Which circumstances can cause the "pre-defined" results? To answer these questions, RC methodologies will be introduced.

Different RC methods can be applied according to the given problem. One approach is that if we cannot find any good results with one method, then we choose another one. The number of analysis can be multiplied not just with the number of chosen methods but with methods' different input parameters range and with seeking and removing outliers. It is not defined when the data are not related to each other. When we use more and more methods with different circumstances, e.g., different parameters and error rates, then we cannot be sure whether a true correlation was found or just a random one. This is true in the case of datarows as well. If we increment the number of datarows, then there will be such a case when we surly find a correlation. Rate *C* shows us the limit number of the datarows or method of analysis. In the case of greater number of *C*, at least two datarows from the set of datarows will surely be correlated.

The second approach means that two (or more) methods produce opposite results. But this is not detected, since we stop at the first method with satisfying result. It is rather typical finding more precise parameters based on the "correlation found" method. There can be two possibilities: (1) two or more methods give the same "correlated" result and (2) one or more methods do not present the same results. In option (1), we can assume that the data items are correlated truly with each other. In option (2), we cannot make a decision. It is possible, that the given methods present the inconsistent results occasionally or they always produce the conflict near to the given parameters and/or data characteristics. There is a specific case in this group, when one method can be inconsistent with itself. It produces different type of results near to given circumstances, e.g., sample size. To measure the random level of this approach, all possible datarows must be produced based on the range of the datarow's values, i.e., between the lowest and the highest values. R shows us the rate how many produced datarows eventuate the same result as the original one.

The third approach is related to the time. The classic approach is that the more data we have, the more precise results we get. But it is a problem if a part of the datarows produces different results than the larger amount of the same datarows. For example, a datarow is measured from start time t_0 to time tt_n, another is measured from start time t_0 to time $t_{n+k}t+k$, and the two data sets make inconsistent result. This is critical since we do not know which time interval we are in the data collection. For this problem, the cross validation can be a solution. If all subsets of the datarow for the time period t_it are not perform the same result, we only find a random model at high probability level. If they fulfill the "same result" condition, we find a true model likely. In this case, *S* shows us the rate of number of the element of the "correlated" and the "non-correlated" subsets.

The third method has another interpretation. In general, we have huge amount of data sets, and we would like to get correlation between them. In other words, we define some parameters, which were or will be measured, and we analyze these datarows and create a model. We measure these data items further [time $t_{n+k}t+k$]. If the new result is not the same as the previous one, we found a random model likely. The reason could be that a hidden parameter was missed from the given parameters list at the first step. It is possible, that the values of these hidden parameters changed without our notice, and the model collapses. This kind of RC is difficult to predict.

Applying CReMIT in Space Domain

Earth Observation data is available from different sources with different resolutions. ESA's Copernicus portal offers free Sentinel images with a resolution of 10 meter covering the Earth approximately every 6 day, while an American company, Planetlabs offers images for its customers with 3, 5 and 0.72 meter resolution covering the Earth every day. These data can be combined with information arriving from social networks (e.g., Twitter, Facebook, Instagram or Google searches). Just as example in the health domain: Earth observation satellite, global positioning satellite and weather satellite data can be used to forecast changes in micro climate while crowd searching can help to identify different diseases at specified areas. By combining these two data, fertilized areas can be identified and

forecasted even in a personalized way (e.g., "Due to the actual water level and weather forecast, the number of the mosquitos can be increased, and in your city many people started to google the symptom of malaria, so it is better to stay at home.") Typical use cases for such combinations are in the domain of disaster management, land management, global health and even more. EO satellites can contribute to 12 of the 17 sustainable development goals of the United Nations including no poverty, zero hunger, climate action, live below water and live on land. But for such combinations, we need to use Big Data approaches.

The importance of Big Data in Space is dynamically increasing in both business and scientific fields. As we saw before, there are many solutions and researches related to satellite data. However, these researches deal with image processing, and the analysis phase of the research is related to the images, and metadata of the images. Nevertheless, there are less researches focusing on seeking correlations based on related numerical values.

CReMIT produces new datarows derived from the original data set. With these new derived datarows, the possibilities of analyses are increased, therefore new correlations can be found including time-shifted and seasonality correlations. However, according to RC theory, the increasing number of analysis possibilities can lead us to the random models. Therefore, CReMIT and RC methodologies supplement each other and work as a framework.

In this framework, there are three main steps. First, the derived time series are created by CReMIT based on the user defined parameters, SP, I and I. The second step is the selection and execution of the method of analysis. Since CReMIT does not depend on the further applied analysis methodology, all kinds of analysis methods can be applied on the CReMIT derived time series, from the simple linear regressions through the different non-linear methods ending with statistical tests. Therefore, the whole framework, included CReMIT and different relationships seeker methods can be used for Earth Observation data as well. After the first two steps (CReMIT and analysis), RC is applied to determine the reliability of the given results. All three RC approaches can be used in the third step of the framework.

The first approach is related to the number of datarows or method of analysis. In practice, we have a datarow *A* and if this data row does not correlate with another one, then more datarows are used to get some kind of connection related to *A*. The question is how many datarows are needed to find a certain correlation. We seek that number of datarows, in which case a correlation will surely be found. There is a rule of thumb stating that from 2 in 10 variables (as datarows) correlate at high level of probability, however we cannot find any proof related to that statement. It rather is a statement based on experiences. The calculation process can be different, depending on the given method of analysis. However, the general process can be followed in all cases. We generate candidates after each other and during in one iteration we compare the current generated candidates with all subsets' all candidates. If we find a correlation between the current candidate and either of the candidates, then the current candidate goes to the proper subset. Otherwise, a new subset is created with one element, i.e., with the current candidate. It is true for each subset that every candidate in the given subset is correlated with each other. Let us denote by *C* is the number of subsets. *C* shows us that how many datasets must be measured during the research to get a correlation with at least two datasets for sure. Based on the value of *C*, we have three possible judgements:

- C is high. Based on the given RC parameters, it must be lots of datasets to get a correlation with high probability. This is the best result, because the chance of RC is low.
- C is fair. The RC impact factor is medium.
- C is low. This is the worst case. Relatively few datasets can produce good correlation.

To determine rate *R*, all data value possibilities must be calculated. The calculation process is based on the lowest and the highest data values related to the given datarow. Then, all possible datarows are produced, which the researchers can measure at all. In this set, one member is the original datarow. The given analysis method is

executed in the matter of all members of the set. *R* shows us that rate between the "correlated" and the "non-correlated" notations. The *R* rate must be determined for all member of the *tsw* vector.

In the case of the third RC approach, all subsets of the given data items are produced. For all subsets, the given analysis method is performed. Rate *S* shows us that how many subsets eventuate "correlated" result compare to that subsets which do not. We execute the given analysis method starting with the first *k* elements of the given datarow. If the result has a "correlated" notation, the counter is increased by 1. Then we attached the next element to the previous subset and perform the analysis method again and the result is noted again. We continue this algorithm until we have data items, i.e. subsets are generated and checked with the given analysis method. At the end, the rate *S* can be determined, which can show that how stable the whole datarow related to the given method. If *S* is relatively high, then it means that the datarow changes very often the "correlated" and "non-correlated" notations and it is not stable. It follows that if we continue measuring data, the next data items can cause opposite result with high probability. In other words, the datarow with the given method cause "correlated" notation, but sometimes not, and this is not an exact result.

Summary

The satellites-based Earth Observation (EO) generates typical Big Data datasets on the basis of data volume, data type and structure. The data, value added services and Big Data analytics from EO will represent a more than $50 billion opportunity over the next decade. This market will grow rapidly because on the one hand the price and cost of satellites will decrease, and on the other hand there will be more and more products and applications connected to this area. There are many different platforms to store and handle this data and they provide services for different domains. However, these platforms usually give services, which ensure the data achievement for the connected applications. The processing of the data is a domain specific task and it demands Big Data techniques.

In this paper, two self-developed methodologies were introduced to process Big Data, both can be applied on space data. The CReMIT method combines the solution of moving intervals and evolutionary techniques to grow the sphere of the examined variables. The input of CReMIT must be a periodical time series, and it systematically generates a huge number of secondary time series based on user defined parameters. These created, new time series can be the input of different analysis methods. Random Correlations theory was also introduced in this paper. With increased number of analyzing possibilities, that environment can be created, where correlations can be born just randomly, and this property of the analysis environment is hidden from the scientists as well.

The application of these methodologies in the space domain makes possible to perform special data analysis and determine the reliability of the results and relationships. Our aim is to develop a general platform to process and analyze EO data, and then to store and visualize the results. The main process methodologies will be the CReMIT and RC, but other analysis methods will be implemented as well.

Acknowledgement

The research has been supported by the UNKP 17-4-III New National Excellence Program of the Ministry of Human Capacities.

References

[1] Arviset, C., Salgado, J., González, J., Gutierrez-Sanchez, R., Segovia, J. C., Durán, J., Hernandez, J., Merín, B., Nieto, S., OMullane, W., Lammers, U. (2016): Big Data, Big Data Challenges And New Paradigm For The Gaia Archive, Proc. of the 2016 conference on Big Data from Space (BiDS'16), pp. 9-12

[2] Zikopoulos, P. C., Eaton, C., deRoos, D., Deutch, T., Lapis, G. (2011): Understanding Big Data, McGraw-Hill Osborne Media, 2011

[3] Pödör, Z., Edelényi, M., Jereb, L. (2014): Systematic Analysis Of Time Series – CReMIT, Infocommunications Journal 6 (1), pp. 16-21

[4] Bencsik, G., Bacsárdi, L. (2016): Novel Methods For Analyzing Random Effects On ANOVA And

Regression Techniques, Advances in Intelligent Systems and Computing 416, Springer, pp. 499-509
[5] Bencsik, G. (2016): Decision Support And Its Relationship With The Random Correlation Phenomenon," Ph.D. Dissertation
[6] NSR's Satellite Based Earth Observations, at: http://www.nsr.com/ (Last visited: Jan 10, 2018)
[7] Esch, T., Asamer, H., Balhar, J., Boettcher, M., Boissier, E., Hirner, A., Mathot, E., Marconcini, M., Metz, A., Permana, H., Soukup, T., Uereyen, S., Svaton, V., Zeidler, J. (2017): Monitoring urbanization with Big Data from space - the Urban Thematic Exploitation Platform, Proc. of the 2017 conference on Big Data from Space (BiDS'17), pp. 243-246
[8] Häme, T., Tergujeff, R., Rauste, Y., Farquhar, C., van Zetten, P., Kershaw, P., de Groof, A., Hämäläinen, J., van Bemmelen, J., Seifert, F. M. (2017): Forestry-Tep responds to user needs for sentinel data value adding in cloud, Proc. of the 2017 conference on Big Data from Space (BiDS'17), pp. 239-242
[9] Albani, S., Lazzarini, M., Nunes, P., Angiuli, E. (2017): A platform for management and exploitation of big geospatial data in space and security domain, Proc. of the 2017 conference on Big Data from Space (BiDS'17), pp. 275-278
[10] Soille, P., Burger, A., De Marchi, D., Hasenohr, P., Kempeneers, P., Rodriguez, A. R. D., Syrris, V., Vasilev, V. (2017): The JRC Earth Observation Data and Processing Platform, Proc. of the 2017 conference on Big Data from Space (BiDS'17), pp. 271-274
[11] Biondi, F., Waikul, K. (2004): DENDROCLIM2002: A C++ program for statistical calibration of climate signals in tree-ring chronologies, Comp Geosci 30, pp 303–311

Measuring Organizational Information Security Awareness Levels Supported by a Maturity Model

GÁBOR TARJÁN
Corvinus University Budapest, PhD student
eMail: Gabor.Tarjan@magicom.com

ABSTRACT

With reference to a concept of Information Security Awareness (ISA) this article talks about the importance of measuring ISA as a focal point for an improved competitiveness of organizations. The measurement possibilities are limited by some ethical aspects and some measurement scale related challenges are also existing. Providing a balanced and selected picture of maturity models we describe a theoretical maturity model (MM) for measuring ISA in organizations. The presented model in this paper is waiting for tests and validations but as an initial step for further studies can be considered. Based on this study we have a solid basis for modelling ISA strengths in various organizations. The value of the elaborated ISA MM will be tested and validated in the next phase of the research but it is important to initiate a discussion about the limitations of the ISA MM in an early phase of research.

Introduction

Information Security Awareness (ISA) is a hot topic in literature, several researchers (Lebek [1], Parsons [2], Nemeslaki [3], Bulgurcu [4], Sasvári [5], Maqousi [6], Siponen [7], Molnár [8]) investigated the related challenges. Each of them provided a definition of ISA from his/her own aspect, but some inconsistency is also observed in wording and in the usage of the established concept. This was the main reason for starting and completing a comprehensive literature review to find an appropriate definition of ISA for a further research.

Nowadays ISA is an absolutely key-concept from the viewpoint of compliance and competitiveness for the profit sector in international business context. Understanding the importance of knowing the ISA strength level in an organization, it is obvious to ask for an elaborated measurement model for ISA.

The aim of this article is to elaborate and publish a unique maturity model for ISA providing a background tool for auditors and assessors facing audit challenges related ISA maturity measurements.

The first section of this paper gives a definition on ISA, which can be a base for our modelling activities. The second chapter answers the important question: Why is important to know, assess and classify organizations based on their level of ISA.

The next section discusses some challenges of measuring ISA in organizations and the following part provides an answer on our measurement challenge: what can be the role of maturity models in this situation. This chapter gives an overview about maturity models (MM) used by different industry sectors and interested parties. This part of the paper also describes the limitations of our measurement efforts.

The next chapter describes an ISA MM which can be used for assessing ISA on organizational level. The last part of the paper talks about conclusions and research questions for future studies.

An ISA concept

Based on a previous article [9] we formed a definition on ISA as follows: ISA is a knowledge and attitude of interested parties of an organization on the protection of information assets owned or managed by the organization. We stated that there are some important layers of this definition:

- ISA talks not only about managers and employees but covers a wide range of interested parties who can have influence on the ISA status in an organization (i.e. we expect some ISA from our clients also in the finance sector, because the followed good practice has a great impact on the security status of a financial institution – see password management and safe PIN usage by card holders)
- *Knowledge:* Knowing the rules, procedures and instructions related to ISA is crucial but itself this type of knowledge doesn't provide active defence on information assets. In this context knowledge involves those skills also, which provide the ability for completing actions required by an existing control in the organization.
- *Attitude:* This means that there is an active and positive approach on security related controls and countermeasures. The people not just understand what to do and why is it right, but they are actively involved into preventive and corrective actions. They report suspicious activities observed, they are involved into backup and recovery activities, they follow the rules and actively advise each other when there are unexpected challenges.
- Owned or managed information assets: The ownership of the information is important but not the only one factor determining the organizational behaviour. This new era of data processing very often creates such a situation when the processor is responsible for the information security related issues but the data is not owned by itself (see cloud technologies or any agencies who are responsible for data processing). These special cases have serious influence on ISA programs and campaigns completed by the referred organizations.

Based on this layered definition we can have one basic question: How are we able to measure the level of ISA in a certain organization?

Importance of the Measurement

The measurement in science is very important, but working as an IS auditor we meet this need from managers as well. Managers express their interest in measurements in the field of ISA also. This need is formed in the following questions:

- What is the level of ISA in a certain organization/unit?
- Are there any changes in the ISA status since the last audit event (increasing or decreasing ISA in a time-period)?

- Are there any KPIs on ISA and what indicates these numbers?
- Is the visualisation of ISA (i.e. dashboards) possible?
- What are our ISA "results" comparing to our competitors or industry leaders?
- Is there any relation between organizational performance and ISA level in an organization?

The managers, who are aware of IS related risks, are interested in the status of ISA in their own organization. They need to know whether this status is better or worse than it was in a previous period. These managers want to know the nature of changes and why increasing or decreasing ISA. As managers, they would like to simplify the question by using key performance indicators (KPIs) and if it is possible they would like to see a simple dashboard.

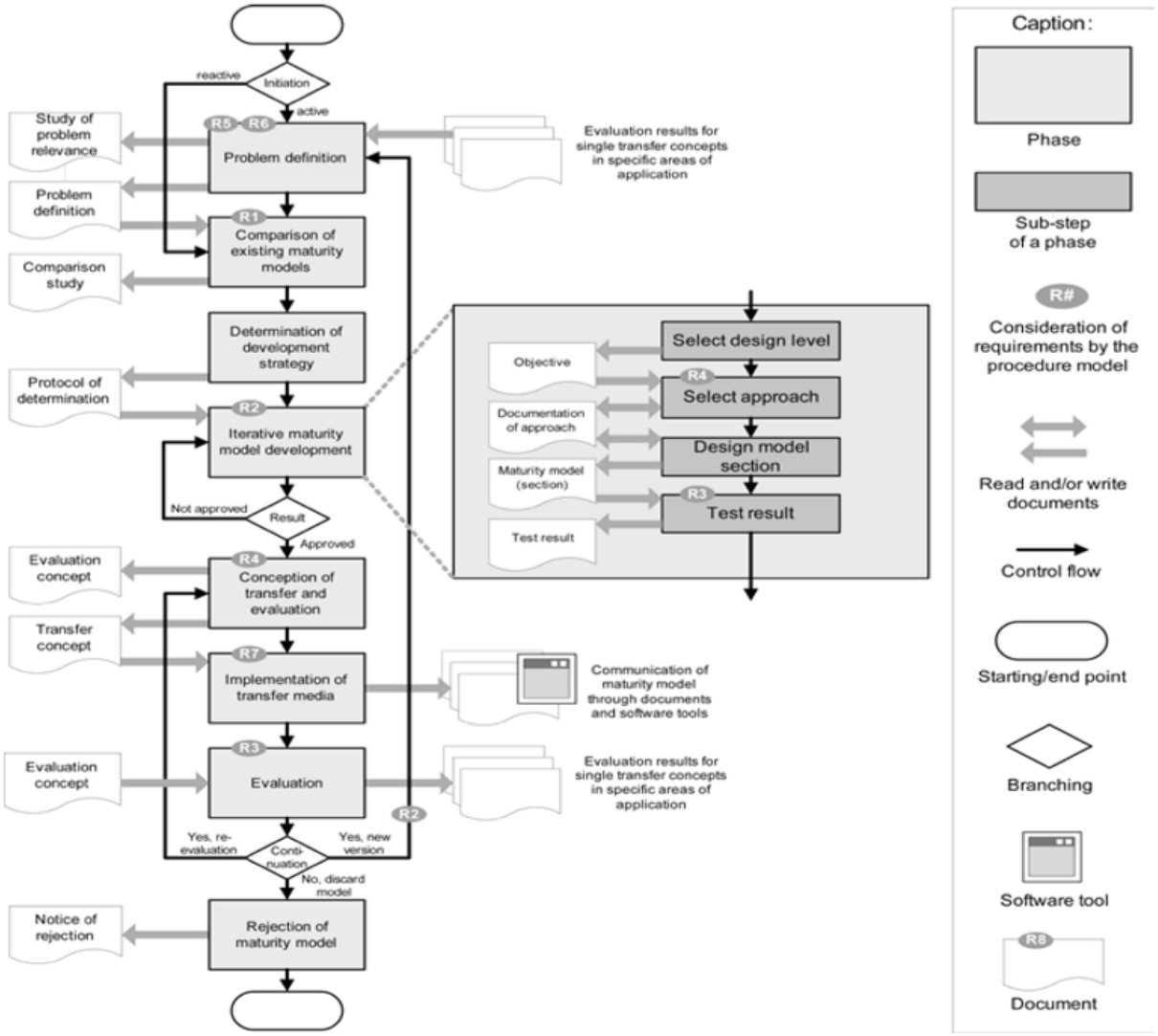


Figure 1. Procedure model for developing maturity models [14]

How to measure ISA

Avoiding personal conflicts generated by audit events the professional auditor tries to provide a realistic picture on ISA on the level of organization and not on the level of individuals. This person-neutral approach has some benefits for every parties:

- The individual will be not claimed or punished based on IS audit statements.
- The auditor will not hurt any ethical standards of auditing
- The company will receive a holistic picture on ISA status as a whole (as an organization)
- The manager will get a clear picture on missing or malfunctioning controls and/or areas for improvement

Based on these identified common benefits for all interested parties we need to find a way how to deploy useful observations on ISA status and areas for improvement. The primary tool for this deployment called "maturity model" (MM), which is widely used by different industry sectors, branches and science fields. MMs are existing in numerous areas of management science i.e.

- in Project Management (Organizational Project Management MM – OPM3) [10],
- in Business Analytics (IBM Big Data MM) [11],
- in Software Development (Capability MM) [12],
- in Human Management (People Capability MM) [13],

Naturally this list is not a full and comprehensive list of existing MMs, but indicates how well covered is this topic by the management science literature. The Information Technology is extremely covered by MMs. Demonstrating this proliferation, Becker [14] and his co-authors provide a procedure model for developing maturity models as it is shown by Figure 1. [14].

About the Maturity Models

What is a maturity model? As Gabor Klimko [15] says "maturity models describe the development of an entity over time". In our case we would like to use an appropriate model for describing the ISA level changes over time in an organization. There are some general considerations on maturity models [15]:

- The development stages of the entity are defined with a limited number of maturity levels (generally four to six).
- The maturity levels are described by some requirements which the entity needs to achieve on a certain level.
- The maturity levels are sequentially ordered (Ordinal Scale of Measurement!) from an initial level up to an ending level (the highest level represents the perfection).
- During a development process, the entity is moving forwards (or backwards) from one level to the next one and no levels can be left out – with other words: the entity is not able to jump or move from a defined level to a faraway existing level without touching each intermediate level.

These considerations will be very important when we try to develop an appropriate maturity model for ISA. Defining a maturity model we can't avoid touch some measurement related considerations either. The Science of Statistics knows four basic scales of measurement. The measurement scales are classified by some properties of measurement:

- Identity: Each value on the measurement scale has a unique meaning.
- Magnitude: Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- Equal Intervals: Scale units along the scale are equal to each another.
- A minimum value of zero: The scale has a true zero point, below which no values exist.
- The four basic scales of measurement:
- Nominal Scale of Measurement: This scale only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude. *Because of these reasons the nominal scale is not an option for measuring ISA in organizations.*
- Ordinal Scale of Measurement: The ordinal scale has the property of both identity and

magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale. *The maturity models are typically representing this type of scales and can be used for our measurement purpose, therefore we focus on that type of modelling.*

- Interval Scale of Measurement: The interval scale of measurement has the properties of identity, magnitude, and equal intervals. With an interval scale, you know not only whether different values are bigger or smaller, you also know how much bigger or smaller they are. *In the case of maturity models, you can't use this character because the question "how many times bigger is the ISA in company A than in company B?" has no real meaning or sense.*
- Ratio Scale of Measurement: The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and a minimum value of zero. *This level of measurement is not available in our case, when we would like to express the strength of IS practice (ISA) in an organization.*

The ordinal scale is the second on the list in terms of power of measurement. The simplest ordinal scale is a ranking. The maturity models are using this type of measurement. In this case there isn't objective distance between any two points on the subjective scale. The ordinal scale only lets Us interpret gross order and not the relative positional distances. From a research view the ordinal data allow Us to use non-parametric statistics. These includes median and mode, rank order correlation and non-parametric analysis of variance.

Our statistical limitations are defined and during the validation process of a maturity model we can use only these abovementioned statistics.

The word of maturity models (MM) is very rich and practically unlimited therefore we do not want to provide a full picture of MMs, but we highlight some of them, which are very useful for establishing our ISA MM. The following MMs gave Us some ideas for improving our own MM of ISA:

The ITIL Maturity Model

The ITIL Maturity Model [8] is a scale of six levels of maturity focusing on processes and functions. These maturity level definitions are aligned with COBIT which is also referred in this section below. There are two key concepts, definitions, used by ITIL [16]:

- Definition: process = A structured set of activities designed to accomplish a specific objective. A process takes one or more defined inputs and turns them into defined outputs.
- Definition: function = A team or group of people and the tools or other resources they use to carry out one or more processes or activities – for example, the service desk.

The maturity level definitions of ITIL Maturity Model:

- *Level 0* (absence/chaos): Processes or functions are ad hoc, disorganized or chaotic.
- *Level 1* (initial/reactive): Processes or functions follow a regular pattern but no process or function governance exists.
- *Level 2* (repeatable/active): The processes or function has been recognized and procedures have been standardized, documented and communicated through training.
- *Level 3* (defined/proactive): The activities are appropriately resourced, although occasionally, and in unusual circumstances, may be inadequate.
- *Level 4* (managed/pre-emptive): The process or function and the associated activities are robust and rarely fail to perform as planned.
- *Level 5* (optimized): All activities are subject to management control, governance and leadership.

This ITIL MM provides some ideas for creating our own MM for ISA:

- The lowest level represents the full absence of maturity, therefore our initial level also will represent the total lack of ISA.
- The highest level of ITIL MM talks about management control, governance and leadership, therefore in our own model we need to build in these concepts to the relevant highest level.
- Maximum 5-6 levels can be defined in an appropriate MM for ISA.

The COBIT 5

The most significant methodological product of ISACA (Information Systems Audit and Control Association) is the COBIT (Control Objectives for Information and Related Technologies) as a good-practice framework. The COBIT 5 [17], the latest issue, defines an IT Governance Framework, in which has an important role for a Process Capability Model. The Model classifies processes into six levels:

- *Level 0:* Incomplete process. The process is not placed or it cannot reach its objective. This level the process has no objective to achieve. For this reason, this level has no attribute.
- *Level 1:* Performed process. The process is in place and achieves its own purpose. This level has only "Process Performance" as process attribute.
- *Level 2:* Managed process. The process is implemented following a series of activities such as planning, monitoring and adjusting activities. The outcomes are established, controlled and maintained. This level has "Performance Management" and "Work Product Management" as process attributes.
- *Level 3:* Established process. The previous level is now implemented following a defined process that allows the achievement of the process outcomes. This level has "Process Definition" and "Process Deployment" as process attributes.
- *Level 4:* Predictable process. This level implements processes within a defined boundary that allows the achievement of the processes outcomes. This level has "Process Management" and "Process Control" as process attributes.
- *Level 5:* Optimising process. This level implements processes in the way that makes it possible to achieve relevant, current and projected business goals. This level has "Process Innovation" and "Process Optimisation" as process attributes.

As an important rule, in COBIT 5 to achieve a given level of capability, the previous level has to be completely achieved. This rule meets the principle of MMs, that the entity (in this case the process maturity) is moving forwards (or backwards) from one level to the next one and no levels can be left out.Pasquini and Gallé [18] say this improved model asks for non-subjective assessment and evidences and as a result, the reliability and repeatability of process capability assessment activities and evaluations have been improved, reducing disagreements on assessments.

The COBIT 5 Process MM verifies what we already learned from the ITIL MM. Nearing the ISA and its already published MMs we met some models, which are addressing awareness related issues:

The Capability Model (ISACM)

The ISACM was published by Poepjes and Lane [19]. They connected ISO/IEC 27002 with the theories of situation awareness. They had a control based approach and examined ISO/IEC 27002 from that aspect. They identified three dimensions of awareness influenced by controls in the referred standard:

- Awareness Importance: how important is the awareness in the success of the correct functioning of a process or control,
- Awareness Capability: how capable is a person when faced with a decision
- Awareness Risk: This is a gap that results from the required amount of awareness (Importance) being greater than that is displayed (Capability)

These dimensions are connected with the identified controls required by the ISO/IEC 27002 standard. As a part of this model the stakeholder groups are also identified and connected to the controls. The MM is summarised as it is shown by Figure 2. [19 p7].

The ISACM defines three important dimensions of awareness: importance, capability and risk. It also creates groups of internal interested parties (IT staff, senior management and end users) but this grouping demonstrates the model limitations as well: It talks only from the aspect of IT (and ISA is not only about IT) and the external interested parties are not considered in the model. The ISACM also provides some ideas for creating our own MM for ISA:

- Some elements (some identified controls) from ISO 27000 family of standards can be used for

creating an "inventory of controls" related to the defined levels of ISA MM.

- Awareness has more dimensions which we need to consider during our model creation.
- The stakeholder group's coverage is also important in a MM for ISA.

After these shortly presented institutional MMs, we need to focus on individuals who are also assessed in MMs as it follows

Information Security Awareness Capability Model

ISO/IEC 27002 Controls Standard	Stakeholder Group	Awareness Importance							Awareness Capability							Awareness Risk
		Importance (influence) that awareness provides to the controls for each stakeholder group. How much awareness is required?							Level of Awareness being displayed by each Stakeholder category.							Highlights gap in required awareness - Interface with Risk Assessment matrix
ISO/IEC 27002 list of controls		None		Slightly	Moderate		Very	Extremely	None		Slightly	Moderate		High	Expert	Overall Rating
5 Security policy																
	Objective: To provide management direction and support for information security in accordance with business requirements and relevant laws and regulations.															
5.1 Information security policy	IT Staff	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Senior Management	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	End Users	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
6 Organization of information security																
	Objective: To manage information security within the organization.															
6.1 Internal organization	IT Staff	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Senior Management	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	End Users	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Objective: To maintain the security of the organization's information and information processing facilities that are accessed, processed, communicated to, or managed by external parties.															
6.2 External parties	IT Staff	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Senior Management	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	End Users	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
7 Asset management																
	Objective: To achieve and maintain appropriate protection of organizational assets.															
7.1 Responsibility for assets	IT Staff	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Senior Management	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	End Users	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Objective: To ensure that information receives an appropriate level of protection.															
7.2 Information classification	IT Staff	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	Senior Management	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None
	End Users	1	2	3	4	5	6	7	1	2	3	4	5	6	7	High/Medium/Low/None

Figure 2. The Information Security Awareness Capability Model (ISACM) [19]

The User Awareness MM

The User Awareness Maturity Model (UAMM) is presented by Steve Kruse and Bill Pankey. They didn't publish the UAMM in specific scientific papers but their presentation is available on the Internet [20]. They assess IT users in a five grade UAMM:

- *Grade 1 - Blissfully unaware:* Uses any capability provided them little recognition or acceptance of most information security threats. At this level, prevalent view is that information security is a property of IT systems and largely a matter of architecture and configuration. Security largely independent of user behaviour.
- *Grade 2 - Consciously incompetent:* Avoids behaviour believed to 'risky', even if that results in some productivity loss
- *Grade 3 – Compliant:* Aware of risks identified in company policy Will take action identified in company security policy
- *Grade 4 - Risk aware:* Considers information security risk in performance of company duties, but unsure of appropriate action; sometime will report incidents
- *Grade 5 - Competent & Practiced:* Expects to manage security risk (recognize and mitigate) when performing duties.

Their UAMM uses two dimensions for putting people into the appropriate grade of maturity as it is in [20].

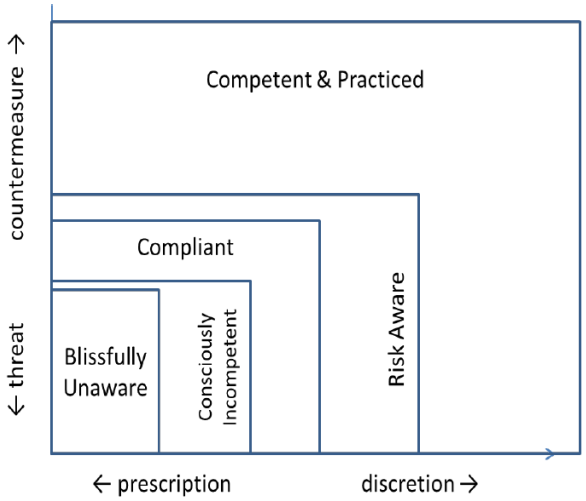


Figure 3. Underlying Maturity Factors [12]

The Figure 3. shows that

- horizontally we can assess user's behaviour on their level of discretion and we can allow more flexibility for users as their maturity increases,
- vertically we can assure higher responsibility in risk management as maturity increases.

The UAMM defines maturity levels related to people and its approach strengthens our commitment to the usage of more dimensions in modelling ISA maturity. The wording of the authors also sends the message that they are talking just on IT related staff ("IT user") when they think on people. From our aspect ISA is not related only those people who are working with IT assets, equipment. Naturally the author's "two dimensions" approach really support Us creating our own MM for ISA.

The SANS Institute – SAMM

The SANS Institute Awareness Maturity Model was published on 22 May 2012 in a security awareness blog by Lance Spitzner. [21] The original model provided a "five-grade" approach as it is presented on this picture [21]. The model in brief:

- *Level 1: No Security Awareness Program* "There is no awareness program, there is no attempt to train and educate the organization. As a result, people do not know or understand organizational policies and procedures, do not realize they are a target, and are highly vulnerable to most human based attacks."

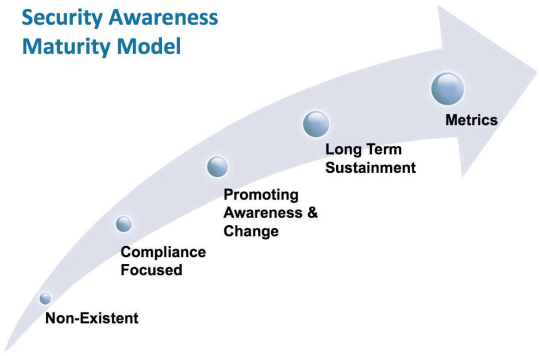


Figure 4. Security Awareness Maturity Model [13]

- *Level 2: Compliance Focused* "This is an awareness program designed primarily to meet specific compliance or audit requirements. Training is limited to annual or ad-hoc basis, such as an onsite presentation once a year or quarterly newsletters. There is no attempt to change behaviour. As a result, employees are unsure of organizational policies, their role in protecting their organization's information assets and how to prevent, identify or report a security incident."
- *Level 3: Promoting Awareness & Change* "On this level, the goal is to have an impact and change behaviours, to reduce risk in the organization. This step is far harder than the first two, and often why you do not see most organizations reach this level. Instead of just ad hoc materials distributed at random times, the awareness the program identifies the training topics that have the greatest impact in supporting the organization's mission and focuses on those key topics. In addition, program goes beyond just annual training and includes continual reinforcement throughout the year. Content is then communicated in an engaging and positive manner that encourages behaviour change at work, home and while traveling. As a result, employees, contracts and staff are aware the organization policies/processes and actively prevent, recognize and report incidents."
- *Level 4: Long term sustainment* "Long term sustainment builds on an existing program that is promoting awareness and change. It adds the processes and resources in place for a long-term

life cycle, including at a minimum an annual review and update of both training content and communication methods. As a result, the program becomes an established part of the organization's culture and is always current and engaging."

- *Level 5: Metrics* "This final level is defined as a security awareness program that has metrics in place to track progress and measure impact. As a result, the program is continuously improving and able to demonstrate return on investment. This is not to say that you cannot use metrics in the previous maturity levels, instead this means we have a formal metrics program." [21]

The same model is presented in the latest SANS Institute Security Awareness Report with some slight changes in naming of grades [22] as shows the following picture:

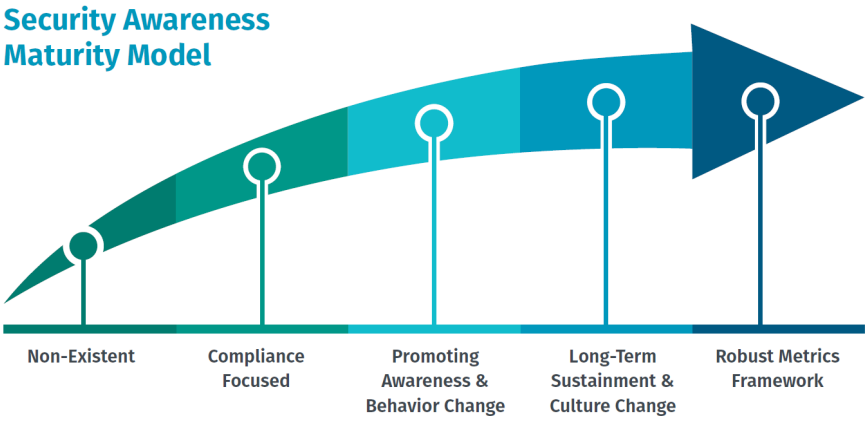


Figure 5. Security Awareness Maturity Model [22]

Originally this model was created for assessing ISA *program* maturity levels but the approach and the layout of this MM is absolutely fit for our purpose: We can use it for creating our own MM of ISA and just we need to add dimensions reflecting to our definition on ISA.

A Proposed Maturity Model

As it was demonstrated in the previous chapters of this paper, there is a significant number of various maturity models which can be applied for measuring ISA. Based on the ISA definition referred in Section 1 (*ISA is a knowledge and attitude of interested parties of an organization on the protection of information assets owned or managed by the organization.)* we just need

- to accept the referred SANS ISA MM [14] as a basis for defining maturity grades
- to add two dimensions (knowledge and attitude) to each defined maturity grade
- to define interested parties involved into each maturity grade
- to make an inventory of controls which can be a proof of existing ISA on each grade
- to add some necessary comments and / or evidences for a common understanding and deployment to each grade

If we completed these abovementioned tasks, our ISA MM is ready for validation. Not forgetting our main aim to provide an auditable, measurable model, we need to define those objective evidences which can be used for assessing ISA maturity in an organization. The below described model is aimed to be fit for this purpose. The proposed model in a visualised form:

As we already stated the "knowledge" dimension involves those skills also, which provide the ability for completing actions required by an existing control in the organization. The "Attitude" dimension talks about the intensity of an active and positive approach on security related controls and countermeasures.

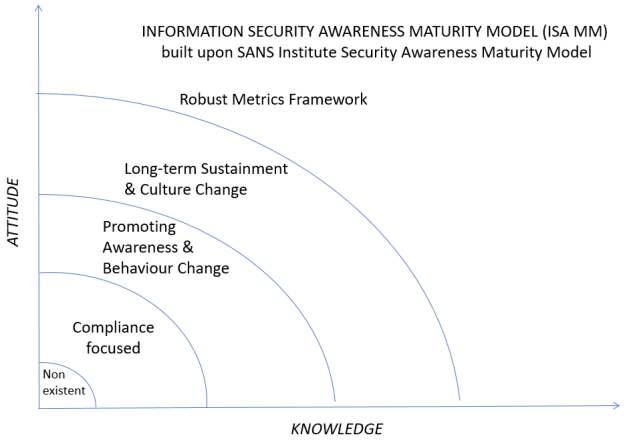


Figure 6. The proposed model for Information Security Awareness Maturity (Own figure)

The detailed model description of ISA MM at organizational level:

Grade 1-Non-Existent: ISA practically doesn't exist.

- Knowledge: Employees have no idea that they are a target, that their actions have a direct impact to the security of the organization, don't know or understand organization policies, and easily fall victim to attacks.
- Attitude: Employees relate neutral or adversely to information security related duties / issues / potential incidents.
- Controls: There aren't supporting controls.
- Stakeholder view: Interested parties are not identified.
- Audit evidences: none

Grade 2-Compliance focused: ISA program already exists but it is designed primarily to meet specific compliance or audit requirements.

- Knowledge: Training is limited to annual or ad-hoc basis.
- Attitude: Employees are unsure of organizational policies and/or their role in protecting their organization's informational assets.
- Controls: Entry and leaving process defined, regular training process defined, internal audits are performed and documented
- Stakeholder view: Customers, Suppliers and the State is considered.
- Audit evidences: Training materials, training records, documented procedure for identification of customer needs, documented procedure for supplier management, documented procedure for initial and regular ISA training, signed NDAs with employees and suppliers, 3rd party audit reports, certificates of compliance issued by customers and/or third parties, risk assessment reports

Grade 3–Promoting Awareness & Behaviour Change: This ISA grade is based on a detailed risk assessment which identifies the topics that have the greatest impact in supporting the organization's mission and the ISA efforts focus on those key topics.

- Knowledge: The training program goes beyond just annual training and includes continual reinforcement throughout the year. The knowledge is tested regularly.
- Attitude: The ISA related content is communicated in an engaging and positive manner that encourages behaviour change at work and at home. As a result, people understand and follow organization policies and actively recognize, prevent, and report incidents.
- Controls: Regular management reviews completed, ISA related projects are completed in a controlled environment
- Stakeholder view: Employees are also considered.
- Audit evidences: List of relevant ISA related topics linked with a detailed risk assessment, management review meeting minutes, ISA project related documents (PID, project plan, action plan, reports etc.), regular management communications on emerging risks, actions, countermeasures and results via e-mail, blog, video etc.

Grade 4–Long-term Sustainment & Culture Change: There is an ISA related program, which has the processes, resources, and leadership support in place for a long-term life cycle, including, at a minimum, an annual review and update of the program. The program and security is an established and updated part of the organization's culture.

- Knowledge: Continually changing learning content with reference to the emerging risks and incidents observed. Knowledge transferred and random tested by unconventional methods.

- Attitude: Strong and positive approach by all interested parties, regulations and policies are actively followed in work, during travel and at home
- Controls: long term planning process and procedure, regular reviews of ISA related learning contents and forms (communication channels)
- Stakeholder view: Every interested party is covered and considered.
- Audit evidences: Program related documentation (set of projects, project and program reports), detailed ISA budget for a longer period (i.e. three years)

Grade 5–Robust Metrics Framework: The ISA program has a robust metrics framework to track progress and measure impact. Consequently, the program is continuously improving and able to demonstrate return on investment.

- Knowledge: Organizational ISA related goals and aims are known by all interested parties
- Attitude: Organizational goals and aims are internalized by every stakeholder and common values exist influencing the daily practice
- Controls: Consistent metrics framework
- Stakeholder view: Nothing to add to the previous grade.
- Audit evidences: documented and traceable KGIs and KPIs, ROI (ROSI) calculations

Some important remarks to the presented ISA maturity model:

- There aren't sharp and rigid borders between grades. In some case the good or bad practices are overlapped and not too easy to decide which grade is appropriate for the assessed organization.
- A certain higher grade always involves the good practices from the lower grade. If you find such state of an organization that the best practices of the highest grade are simultaneously exists and operate with the poorest practice of a lower grade than you need to consider the lower grade as the result of the assessment.
- Although the phrase "metrics framework" is mentioned only at the highest stage, it doesn't imply that the methods of measurement are limited to the last stage of the maturity model. The metrics are an important part of every stage. The highest stage simply reinforces that to truly have a mature program, you must not only be changing behaviour and culture, but have the metrics framework in place to demonstrate that ability of change.

Conclusions, Research Questions

The MM related literature is very rich and it is not too easy to find the appropriate model for measuring ISA. Some elements of existing MMs can be used for a specific one, which is appropriate for showing changes of maturity levels and provides information on moves among levels.

The presented model reflects on the two important dimensions of ISA: Knowledge and attitude also received a balanced role in the elaborated model. This moment the presented model is only theory, but it is suitable for further discussion. The missing validation efforts are the key elements for the next improvement steps.

Validation of the model can be completed via broad based survey among Hungarian companies in the public and private sector. If this validation effort shows some results than an international survey can help in strengthening of the presented ISA MM.

We also need to talk about a missing layer of the model: Which specific controls are operating on each single levels of ISA MM? The existence of specific controls directly indicates the maturity of ISA related activities. The inventory of such controls is the subject of the further study. The results will be presented in a separate paper. Some further research questions (RQ):

- Is the presented ISA MM fit for a repeatable and precise measurement?
- Is the model able to provide evidence on ISA changes by time (i.e. increase or decrease of ISA maturity)?
- Is there any significant difference on ISA maturity between public and private sectors?
- The size of an organization creates any difference on ISA maturity?

- Sensitive industries (like finance and healthcare sectors) provide a higher level of ISA maturity than the low risk industries?
- Can we have valid comparisons among companies on ISA maturity?

The study results will be presented in a separate paper.

References

[1] Benedikt Lebek , Jörg Uffen , Markus Neumann , Bernd Hohler , Michael H. Breitner, (2014) "Information security awareness and behavior: a theory-based literature review", Management Research Review, Vol. 37 Iss: 12, pp. 1049 - 1092

[2] Kathryn Parsons, Agata McCormac, Marcus Butavicius, Malcolm Pattinson, Cate Jerram, (2013) "Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q)", Computers & Security, Vol. 42 pp. 165-176

[3] András Nemeslaki, Peter Sasvári, (2015) „Empirical Analysis of Information Security Awareness in the Business and Public Sectors in Hungary" Central and Eastern European eIDem EIGov Days 2015, Conference Proceedings, pp. 405-418

[4] Burcu Bulgurcu, Hasan Cavusoglu, Izak Benbasat, (2010) „Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness", MIS Quarterly. Vol. 34. Issue 3. pp. 523-548

[5] Péter Sasvári, András Nemeslaki, Wolf Rauch, (2015) „Old Monarchy in the New Cyberspace: Empirical Examination of Information Security Awareness among Austrian and Hungarian Enterprises", AARMS Vol. 14, No. 1, pp. 63-78

[6] Ali Maqousi, Tatiana Balikhina, Michael Mackay (2013) „An effective method for information security awareness raising initiatives", IJCSIT Vol 5, No. 2, pp. 63-72

[7] Mikko T. Siponen, (2000) "A conceptual foundation for organizational information security awareness", Information Management & Computer Security, Vol. 8 Iss 1 pp. 31-41

[8] Molnár Bálint, Kő Andrea (2009): Információrendszerek auditálása; az informatika és az információrendszerek ellenőrzési és irányítási módszerei, Corvinno Kiadó, Budapest, ISBN 978-963-06-7254-2

[9] Gábor Tarján (2017) „Some Conceptual Questions on Information Security Awareness", SEFBIS Journal, No. XI./2017, pp 10-17

[10] Organizational Project Management Maturity Model (OPM3): Knowledge foundation: An American National Standard, ANSI/PMI 08-004-2008 / published by Project Management Institute

[11] Chris Nott (2015) „A maturity model for big data and analytics", IBM, www.ibmbigdatahub.com, http://www.ibmbigdatahub.com/blog/maturity-model-big-data-and-analytics (22.12.2017)

[12] Watts S. Humphrey (1989), „Managing the Software Process", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA 1989

[13] Curtis, B., Hefley, W.E., and Miller, S. (2002), „The People Capability Maturity Model: Guidelines for Improving the Workforce." (ISBN 0-201-60445-0). Reading, MA: Addison Wesley Longman 2002

[14] J. Becker, R. Knackstedt, J. Pöppelbuss (2009), „Developing Maturity Models for IT Management", Business & Information Systems Engineering, June 2009, Volume 1, Issue 3, pp 213-222

[15] Gábor Klimko (2001) „Knowledge Management and Maturity Models: Building Common Understanding", En: The Second European Conference on Knowledge Management. MCIL, Reading, UK, Bled, Slovenia

[16] ITIL Maturity Model, Axelos Global Best Practice, Axelos Limited 2013

[17] ISACA: COBIT Five: A Business Framework for the Governance and Management of Enterprise IT, Rolling Meadows, IL 60008 USA, ISACA 2012

[18] Alex Pasquini, Emidio Galié, (2013) „COBIT 5 and the Process Capability Model. Improvements Provided for IT Governance Process", Proceedings of FIKUSZ '13 Symposium for Young Researchers, pp. 67-76

[19] Robert Poepjes, Michael Lane, (2012) „An Information Security Awareness Capablity Model (ISACM)", Proceedings of the 10th Australian Information Security Management Conference

[20] Steve Kruse, Bill Pankey (2010), „Assessing the Effectiveness of Security Awareness Training", RSA and Tunitas Group, 2010

[21] Lance Spitzner (2012) „Security Awareness Maturity Model" SANS Institute, Security Awareness Blog, 22 May 2012 https://securingthehuman.sans.org/-blog/2012/05/22/security-awareness-maturity-model (22.12.2017)

[22] SANS Securing The Human (2017) – 2017 Security Awareness Report

The Connection between the Production and the Energy Usage in a Smart Factory

ATTILA GLUDOVÁTZ[1] – LÁSZLÓ BACSÁRDI[2]
[1]Lecturer, [2]Associate Professor – [1,2]Institute of Informatics and Economics, University of Sopron
eMails: [1]gludovatz.attila@uni-sopron.hu; [2]bacsardi.laszlo@uni-sopron.hu

ABSTRACT

According to the newest industrial trends, the static and highly centralized networks are replaced by elastic, distributed, many autonomous, but linked components. By decentralizing industrial management, the goal is to achieve a more efficient resource utilization in the factories. The previous wasteful energy consumption cannot be maintained anymore. The process would impose a limit on production, which cannot be tolerated by the producers. The producing machines have built-in sensors, which can detect the disruption of energy input and its usage rate, and they can assign it with the resulting pieces, so the system can deduct the most efficient settings of the operating machines. At a manufacturing company, the managers aim to develop their production-related processes. This development means not only a new method, but a new mentality, with which we are able to identify different problem areas and solve. The main problem identified earlier is related to the wasteful energy use of machines. Therefore, a data collector and analyzer prototype system was implemented. This system and its possibilities are demonstrated in this article.

Introduction

This study demonstrates the challenges and their solutions of the energy management prototype system. In the last months, the system was created for data collecting. These historical and real-time datasets are the basis of the analysis: first, we have cleaned and prepared the dataset, then the analysis of data is continually executed. This analysis helps us in getting the information. With forecasting methods, we can make estimates for the future events. The managers can get the daily reports about the production numbers and the energy use. In addition, we can focus the occurrences of critical events with defining the limit for risky parameters. The goal is to prevent the stoppage of the production or the excessive use of the energy. After we get the results of analysis, we will make decisions.

First, we have focused on the data collecting process. The company's supervisory system gives us the energy usage data of the selected manufacturing machines. We have chosen different machines, from the aspect of utilization of the energy (high, moderate, standard and low). The measurements are done by the data-collector sensors. We have built the communication network between the machines, sensors and the data-centres. In this way, we can keep under control the complete process in real-time. The energy usage of the machines would be constant, but there is no consideration of the various manufacturing at the company. We notice that they produce circa 1500 different types of products. In addition, there is an upper limit of the energy usage a day, on average, that the company should comply with it. So then, we must collect and select the data, which are connected to the production process. For example, produced products' type and quantity or shift data. At the end of this process, we merged the total sum of the fact data (that mentioned above) through the dimension data (time and machine identifiers). The merging progress is happened every 10 minutes. This time parameter can be modified by us at any time. Then, with help of a business intelligence software, we can analyse the time series of the various data. The managers can make decisions now easier than ever. However, we still have challenges with this system, for example, the reaching the fact data faster or the optimization of the analysis methods.

Our general goals are (1) to make the manufacturing more efficient, (2) to reduce their costs and waste. In our study, we will demonstrate a decision support system at an international furniture company.

About one of our motivating factors, a research group analysed the related literature from 1979 to 2014, and they identified 44 scientific articles in the "energy management in industry" subject. None of these studies investigated the Hungarian industry sector or the wood industrial / furniture company [1].

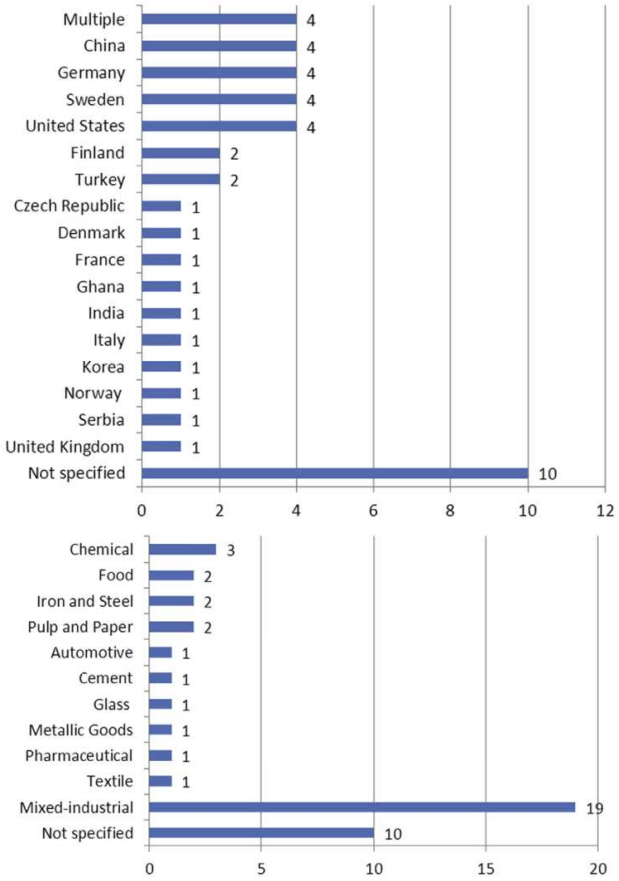


Figure 13. The focuses of studies in the [1] reference's systematic review: (above) geographical, (below) industry sector focus.

The following section introduces the literature review about industrial internet of things and the industrial energy management subject. After this, we demonstrate the design and implementation phases of the energy management prototype system, then the results will be represented. The final section concludes the paper.

Literature review

Industrial Internet of Things (IIoT)

There are a lot of definitions of the IoT (and IIoT), however, in general, we focus their common part, which is the IoT is related to the integration between the physical world and the virtual world of the Internet [2]. In this integrated area, there operate physical objects that want to be able to track, to monitor and to interact among themselves and with the environment. We think, the most user-centric definition for the smart devices is the following: *"Interconnection of sensing and actuating devices providing the ability to share information across platforms through a unified framework, developing a common operating picture for enabling innovative applications. This is achieved by seamless ubiquitous sensing, data analytics and information representation"* [3]. In this system, we can identify IoT elements, which are the smart devices. These ones can be ranked into three levels:

- hardware-made up of sensors and their communication networks,
- middleware – data storage, computing tools and data analytics,
- presentation layer – visualization and interpretation tools for different platforms.

There are several enabling technologies, which are the basis of the IIoT terminology:

- RFID[2] – it is the basis of the tracking the objects process; RFID enables to define labels for identifying the workpieces in the factory [4],
- WSN[3] – it results that the communication will be more efficient, cost-effective and the sensors send the collected data to the distributed or centralized management system for analytics,
- addressing schemes (e.g., IPv6, ZigBee) – they support to uniquely identify and control the production devices through the communication network,
- (new protocols – that define the communication's rules between the elements [5],

[2] Radio Frequency Identification
[3] Wireless Sensor Networks

- (data storage (local, public cloud, private cloud),
- data processing and analytics – these enables smart monitoring and controlling, besides, we can use a lot of new algorithms (e.g. data mining techniques, neutral networks, machine learning methods etc.), the using of these ones results more efficient decision in the manufacturing,
- visualization - it allows the interaction of the user with the system's components; in this level, the data are converting into information and knowledge by the users.

There are a lot of open challenges in the industrial IoT sector, but the decision makers and the experts agree with the end goal is to have plug and play smart devices, which can be applied in any informatics environment for supporting the above mentioned IoT terminology.

Energy consumption optimization

The manufacturing plays one of the important roles within the global economy, besides, industry is the largest consumer of electricity among all end-user sectors. The researches pointed, that the industrial sector consumes about 37% of the world's total delivered energy (more than any other sector) [6]. In a smart factory, the energy savings connects to the effective product life-cycle management (PLM) and the analyzing of the energy consumption data is necessary for the efficient operation. The raw materials, workpieces, components, finished products are the parts of an integrated energy parameters monitoring system, which works in real-time. The data storage and the set of the critical information are available for the management. These things can also be interpreted immediately, and the factory's integrated system will be controllable efficiently. So, the energy consumption status of objects can be monitored during the whole product life-cycle.

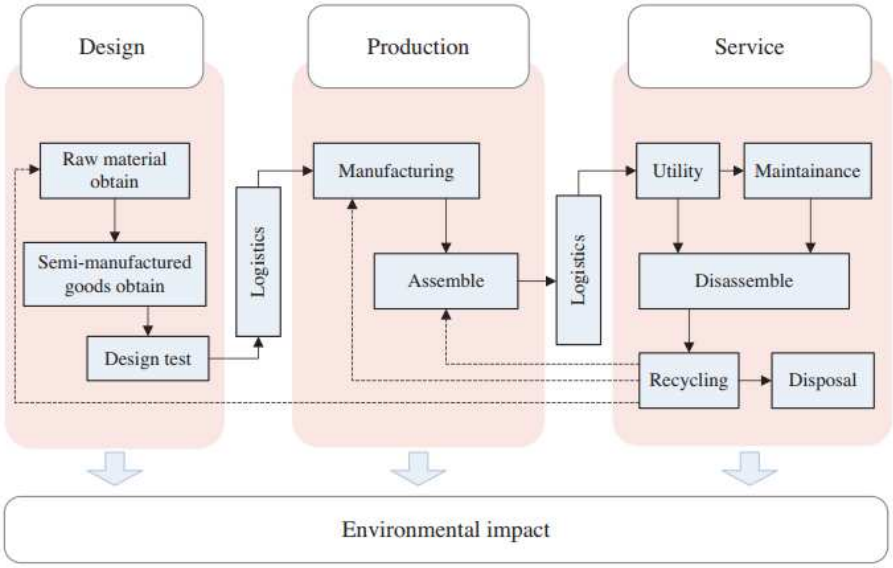


Figure 2. Product life-cycle (Source: the [7] reference)

PLM manages the knowledge intensive processes: the product design, development, manufacturing, distribution and the post-sale services, moreover the product's recycling process too. PLM enables a company to reduce their product-related costs. The PLM has six phases, but the viewpoint of the energy consumption, there are three main parts: design, production and service (see Figure 2). In this study, we are focusing on the *production*, so we will demonstrate this part of the PLM. "*The PLM helps IoT technology to solve problems related to the energy consumption over product life-cycle better*" [7]. The production part makes high percentage of the total energy consumption through the product

life-cycle. Therefore, the potential energy savings is very significant in this field. We would like to realize a huge optimization of the usage of energy. This is our main goal, moreover, we will analyze accumulated production and energy data and find the reasons of energy waste.

The production energy consumption consists of three sides: (1) manufacturing equipment, (2) public facilities (heating, lightning, ventilation etc.), (3) workpiece handling. In our case study, we are focusing on the energy consumption of the manufacturing equipment. We are analyzing the manufacturing equipment's normal working, breakdown losses, equipment's overhaul, the quality tests etc. We have investigated different case studies, that describe several methods for optimizing the energy consumption of a factory, but our solution in a furniture company varies from them (Chakrabarty et al. [8] introduce a significant reduction in energy consumption by changing the state of the devices; Asiedu et al. [9] illustrate that parallel machine tool choice and the energy for "stand-by" affect the energy consumption).

In the following, we will introduce the IoT solutions of the product life-cycle management and its main part, the production. The IoT provides solution (1) to address a lot of elements and their communication, (2) to collect different data, which come from various sources, (3) to get relevant information about the production, (4) to manage the common knowledge related to the operation of the factory, (5) to help to ease the decision-making progress. The main goal is to decrease the energy consumption of manufacturing machines. This process will be improving by the principles and optimized configuration of the equipment.

The IoT is monitoring the status and related parameters of the elements of system, so we can choose a more efficient machine configuration, that also decreases the costs. Thus, we can reduce energy consumption while manufacturing and reduce the idle time (e.g., by switching the machines off – not only stand-by). From the viewpoint of quality management, the waste products can be identified in real-time, thus, we can spare additional costs. Our earlier research connected to this subject: it was a video based IoT solution, which gives information about the quality of the raw materials (timber boards) at a wood industrial company [10].

Key Performance Indicators

Before we implemented the prototype system for energy management, we had defined three characteristics of the system: (1) the energy data standards, (2) the energy performance measures, (3) the optimized energy profiles. The *energy data standards* determine the methods of the data transmission, storage and processing or maybe the sampling periods, levels of data granularity. The *energy performance measures* are like the key performance indicators (KPIs). These parameters measure e.g., that how many products will be complete from specific energy usage, that is a simple ratio parameter, which describes a relationship between an activity and the required energy. In addition, these measures show the results of a wrong development and their critical errors. Every manufacturing sector is different from the viewpoint of energy usage, therefore, there exist different optimization techniques and they use several energy management systems. There are recommendations and references in the industrial sector (e.g., ISO 50001 that is International Standard for Energy Management Systems [11]), but there exist self-made optimizing techniques for efficient working [12][13].

Industrial energy management's goal is to measure, monitor and control the efficiency of energy consumption of the entire production process. That's why, the managers are using key performance indicators, which represent the performance of the processes and they may take effect on the processions (if it's necessary).

There are energy related key performance indicators, which should be independent of industry sector, application etc. A possible list of these KPIs: (1) power and energy consumption, (2) energy costs, (3) energy efficiency and losses (every KPI may be summarize by time / shift / machine / plant etc.) [14].

However, we are focusing on the production data in conjunction with the energy related data. In the following, we are introducing our KPIs, which

are related to the energy management as well as the production too. We show KPIs of a furniture company, there are different types of KPIs (physical, economic, statistical etc.). The physical indicators can be interpreted in the shop floor level, the economic indicators are useful at an aggregated level (e.g., for comparing different plants). In this study, we demonstrate two KPIs in details.

Energy Management System At a furniture company

In this section, we describe an integrated energy management framework and a prototype information system. The energy consumption data come from manufacturing machines (through the building supervisory system), the enterprise data come from the ERP[4] system, then, we combine these ones and make optimizations in the company's operation.

Our research consists the following steps: (1) "state-of-art", literature overview (in the latest sub-sections), (2) design (after a lot of discussions with the engineers), (3) implementation (framework, data acquisition, transmission, storage, processing), (4) data usage and evaluation. In this section, we are focusing on the 2-4 steps.

Design phase

During the design phase, we (1) define the production and energy-based KPIs, (2) make an overview about the production process, (3) select the units, machines, that are resources of our framework, (4) design the database of the energy consumption data, (5) select the relevant data of the production.

Selection of energy related KPIs

First, we defined that want to know about the energy usage at the company. These energy-related data are related to the operation of the factory; thus, we have needed data about the production process. To reach an economic gain, we must join these datasets and investigate the trends and the correlations. Making a rapid survey of the literature's suggestions we made a list about the relevant KPIs at the company [15]. If we would use the abovementioned KPIs, then we have more efficient information about the production.

We must use the most practicable diagrams, which contains the important parts of the results in every single case. For example, if we know the details about the big consumers (units), we must use pie charts for identifying the sub-consumers and their parts of the whole consumptions.

Measurement of energy efficiency

We selected different machines from view of energy consumption and their profiles. In the furniture factory, there are machines for various processes, e.g., squeeze, packaging, shaving, profiling, surface treatment. The machines' energy consumption is dependent on their function, in our selection, there are high and low energy consumers. Aside from the machines, there are other units in our selection for getting more detailed information about the operation of the factory (e.g., ventilation units, compressors etc.). To these units, we built the communication network, which will be describe in the "Implementation phase" section.

Energy usage database

We monitor the energy consumption of units and machines with the help of Supervisory Control and Data Acquisition system (SCADA). This represents the real-time values about the temperature, energy consumption, humidity, brightness and so on.

Furthermore, we configured a scheduled process in the SCADA system. That is the incremental saving into a local database, where we can store, modify or calculate the relevant data. There are four tables in the database. In this structure (as it is illustrated in Figure 4), at the beginning, we stored the metadata of the factories' units (on the left below), the units' parameters' values (in the middle), the parameters' types (e.g., efficient energy use, total efficient power etc. on the right below), and then the measured data (on the left above). After this, we generated a view, which contents the important fields and data for the further analysing steps. (The yellow tags mark the database tables; the green tag marks the database view.)

[4] Enterprise Resource Planning

Table 5. The company's goals, tasks and their required data

Goals	Tasks / Jobs / Details	Required data
Finding and reducing the energy waste sources	– Create reports: details in table, in diagram. Date/time and machine filters. (See the "Case study – First KPI " section.) – Calculating flowrate ratios (e.g., finished products / kWh). – Comparing energy consumption for the same processes. – Configuring the machine learning algorithms to the goal.	Production: products, quantities, machines & configured parameters, shifts data. Energy usage: all available data of machines and units.
Preventing a financial penalty due to break the maximum limit of the energy consumption	– We configured watchdogs, which indicate if the local limits exceeded. The watchdogs will stop the big consumers if the company reaches the global limit. – Create reports: details in table, in diagram. Date/time and machine filters.	All energy consumption data in details.
Increasing energy-efficient maintenance	– Comparing (1) shifts' efficiencies, (2) manufacturing times and other parameters per batches, (3) the production efficiency of given product's different colour versions etc. – Giving suggestions on the optimal working of manufacturing. – It is important in planning of company budget.	Production and energy usage data. In addition, we need not only the energy consumption data, but the thermal efficiency data too.

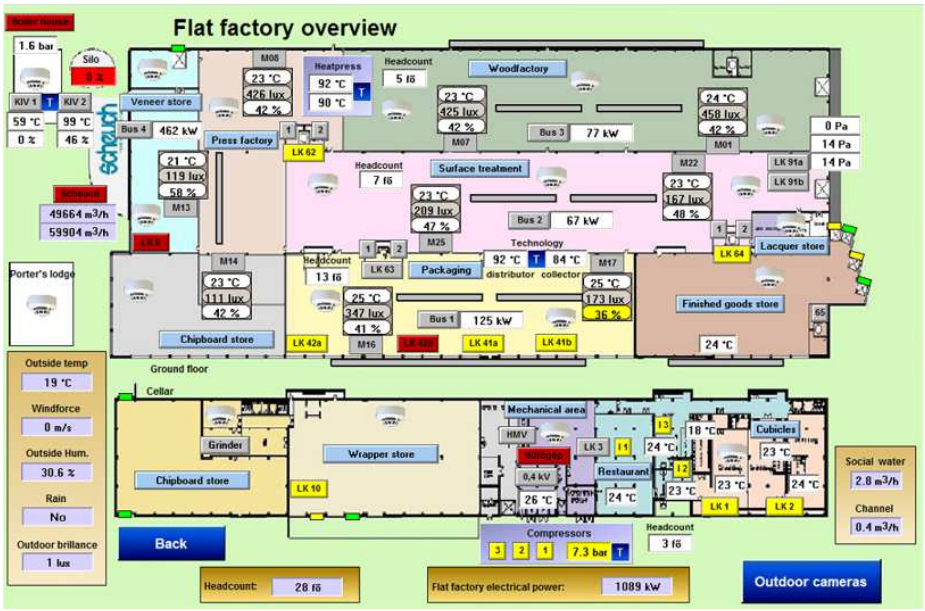


Figure 3. An overview about the operation of a factory at the company
(Source: this is a screenshot from the company's SCADA system)

The relevant data from ERP system

The energy usage data are not enough for us, we need the production data from the ERP system of the company. The relevant data of the production for us: product types (cc. 1500 types), manufacturing quantities, shifts, machines' states, normalized time and cost, Bill Of Material (BOM), Standard Operation Plan (SOP), workshops, work centres, work calendar etc.

Implementation phase

During this phase, we (1) represent an overview about the architecture of the energy management framework, (2) give a brief about our data source system, (3) define the common attributes of our source data.

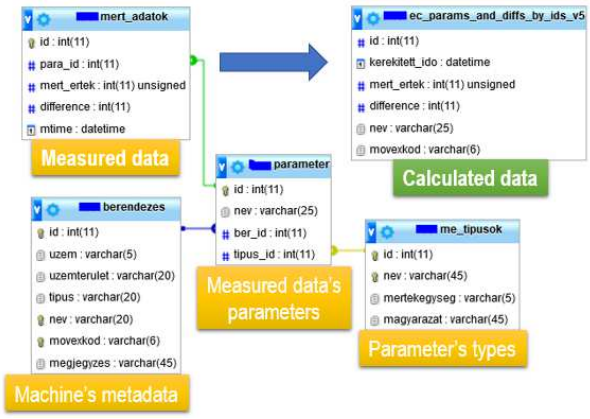


Figure 4. Database diagram about the energy usage in the factories (Source: it is from our database system)

The framework

In this section, we introduce the details of the energy management system from the bottom to the top. At the basis of the framework, there are the selected production lines, machines and other units. We installed the smart meters and sensors to these devices.

At the company, the managers already knew the information about the production's main parameters, so we only extended their communication network for transmission of the energy-related data. We use "Power Meter Series PM9" smart meters, that can connect to an energy management system. In addition, these IoT-compatible smart meters can measure multiple parameters, and they have computational abilities. They connect to the data transmitting network through a Modbus (RS-485) output. (Previously, we defined the list of selected devices (machines, units) for measuring the power quality.)

At the company, there is a gateway that transmits the data to the right system. Thus, the energy consumption data go to the SCADA system, which monitors the data in real-time (or in a trend diagram). The SCADA system's inner database is not available with any management tools, but we can define a routine, a scheduled process or job, with that we are able to make an incremental saving into an external database, which works beside the SCADA system. Thus, we configured a process that saves the energy-related data every 10 minutes. Furthermore, the production data go into the production-related database from the gateway. We can manage these data in the ERP system.

In the energy management system, we use the energy-related and production-related data too. In the *Relationship of the Databasets* section, we show you the basis of joining datasets. After that, we can analyse the datasets together and represent the results via reports. The whole introduced process can be seen in Figure 5.

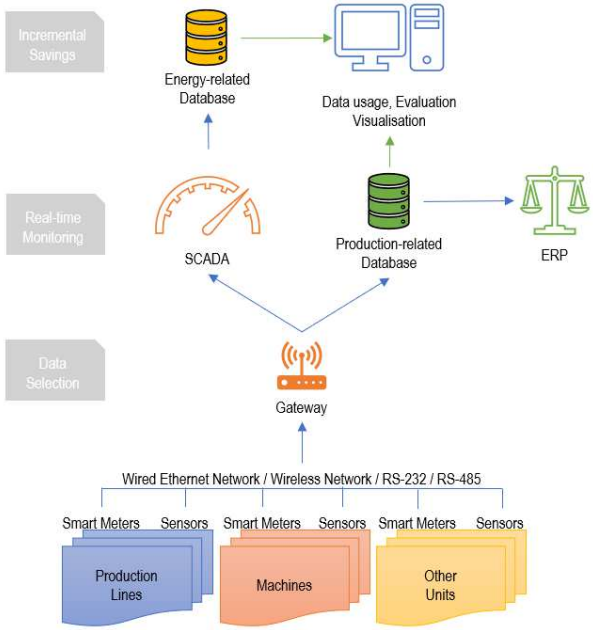


Figure 5. The energy management system's framework (Source: self-constructed diagram)

Energy management system's sources

In our case study, the visualization of energy usage is showed by a SCADA system, its name is VISION X[9]. This system supports the following: data visualization, the connection with operators, controlling, network maintenance, database management and programming possibilities for complex tasks. All of them can be executed in real-time. This system provides us – in a local database system – the energy usage data from units of the production.

The production data are stored in MOVEX ERP system. The corporate processes' data structures in the ERP are the same in the plants. These data are synchronizing continually to the center server of the multinational company.

Relationship of the datasets

We described the relevant energy-related and production-related data earlier, now, we define the relationship between these data: the joins of the datasets are based on the machine ID and the rounded time (in every 10 minutes), these are the common columns in these database tables. We have cc. 2 million records in the energy-related database from last September. We are focusing on this actual fiscal year at the company (from September 2017 till now).

Production data		Common data		Energy usage data	
Product ID	Quantity	Machine ID	Rounded Time	Measured data	Parameter
1	23	4	2017.09.31. 10:00:00	3	1
2	56	5	2017.09.31. 10:10:00	6	1
3	89	6	2017.09.31. 10:20:00	5	1
...	...	...	...	...	...

Figure 6. Sample production-related and energy-related data (Source: self-constructed diagram)

Data usage, evaluation and visualisation

In this section, we introduce a business intelligence tool and investigate two relevant KPIs in details. The most important for us is a physical KPI, the use of this one, we can identify the waste and implement actions to reduce the rate of the low performance in production. Then, we demonstrate an economic KPI, which represents the energy usage in different time dimension at the company. Our other KPIs' results are under review, so these ones won't be demonstrated in this study.

Analysis and visualization

We can join the datasets (production and energy-related) with the mentioned common parameters. After the joining of datasets, we have a lot of possibilities for filtering and analyzing the data. Then we can get the results as tables and as diagrams also. The company has a powerful tool for these steps, this is the QlikView application. QlikView is a business intelligence application used to generate professional queries. The resulting business reports of the queries aid the decision-making process. In the following, you can see two details KPIs' results, which are generated by QlikView. There are situations in which you may get the results slowly, then you can use such an analysis that operates on an in-memory-based database system [16].

Case study – First KPI

Operating states in the factory can be the following: (1) producing = energy consumption + produced finished products (output), (2) waiting or idling time = energy consumption (e.g., machine in "stand-by" mode) + no produced output, (3) stoppage mode = no energy consumption + no produced output. The goal is to reduce time in the second or third state. The changing of the state may be occurred by events (e.g., giving out the raw material or a machine is going wrong). These events can be used for further analysis to result a better energy consumption in real-time. The system gives us possibilities, so we can drill down into details. In the Figure 7, a machine's operation can be shown in detail.

❖ Relation of Production and Energy Usage

Table 6. Several machines' energy consumption data over a week
(Source: the company's production management's sample data)

Machine ID	Machine name	Production Quantity	Used energy WITH manufacturing (kWh)	Used energy WITHOUT manufacturing (kWh)
8970-1	Easy FL	13 176	911	520
8900-1	Szorosor II	0	0	1 367
9000-1	UV-Sor I	30 863	6 240	2 442
9100-1	UV-Sor II	53 981	19 153	9 805
8300-1	Wemhoner II	21 849	1 402	832
…	…	…	…	…
Total (Σ):		*119 869*	*27 700*	*14 966*

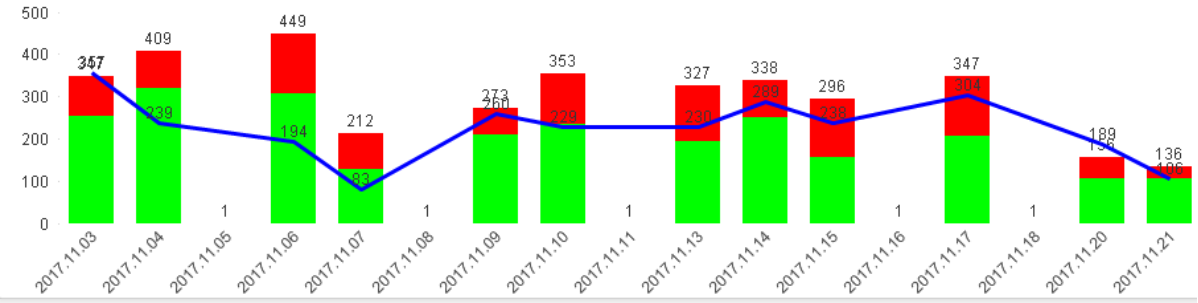
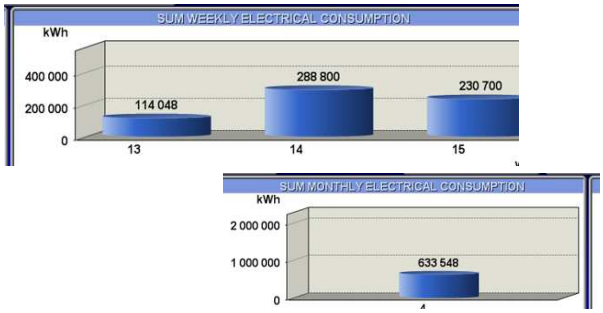


Figure 7. Energy consumption with produced output (green column) and without produced output (red column), the blue series show the produced output on the machine (ID: 8970-1) (Source: self-constructed diagram)

Case study – Second KPI

In this report, we can see the energy consumption over a week, month and the daily details. This report supports the energy planning process at the company. If the managers would like to make more precise decisions, then they must make a connection between these energy consumption values and the prices of a kWh measure. This report represents a company-level summary (it includes two plants and an office).



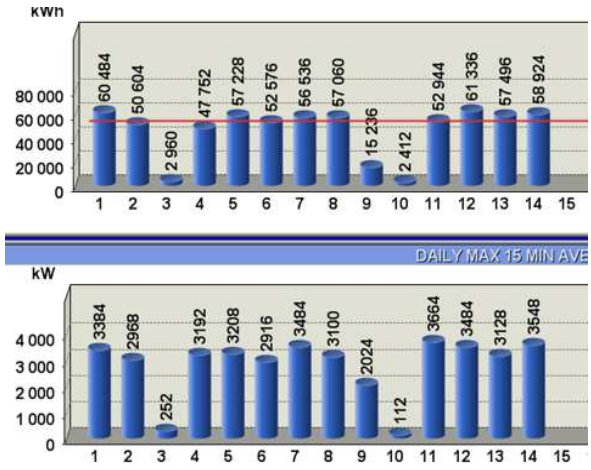


Figure 8. Energy consumption in detail at the company (Source: the company's production management's sample data)

Conclusion

We have introduced an Industry 4.0-related innovation project. After defining the basics of the system, we analysed the operation and the environment of the furniture factory. We have collected the main parameters and KPIs about the factory's working. From the viewpoint of IT development, our framework connects to resources (company's units and other source systems). We designed a database structure and the communication network of our framework as well as created relationships between the energy consumption and the production data. The first results of the data evaluation represent the power of the framework and show the possibilities of our further plans.

Acknowledgment

Supported by the ÚNKP-18-3-3 New National Excellence Program of the Ministry of Human Capacities.

References

[1] M. Schulze, H. Nehler, M. Ottosson, P. Thollander: *"Energy management in industry e a systematic review of previous findings and an integrative conceptual framework"*, Journal of Cleaner Production 112, 2016, pp. 3692-3708.

[2] S. Haller: *"The Things in the Internet of Things"*, SAP Research Center Zurich, 2010.

[3] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami: "Internet of Things (IoT): A vision, architectural elements, and future directions", Future Generation Computer Systems 29, 2013, pp. 1645-1660.

[4] A. Gludovátz, L. Bacsárdi: *"Industry 4.0 projects' background and our experiences at the wood industrial manufactories"*, SEFBIS JOURNAL XI, 2017, pp. 34-41, HU ISSN 1788-2265

[5] Z. Pödör, A. Gludovátz, L. Bacsárdi, I. Erdei, F. N. Janky: *"Industrial IoT techniques and solutions in wood industrial manufactures"*, Infocommunications Journal, Volume IX, Issue 4, December 2017, pp. 24-30.

[6] International Energy Agency (IEA), 2013. International Energy Outlook 2013

[7] F. Tao, Y. Wang, Y. Zuo, H. Yang, M. Zhang: *"Internet of Things in product life-cycle energy management"*, Journal of Industrial Information Integration 1, 2016, pp. 26-39.

[8] V. Swaminathan, K. Chakrabarty: *"Energy-conscious, deterministic I/O device scheduling in hard real-time systems"*, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 22 (7), 2003, pp. 847–858.

[9] R. Schmitt, J.L. Bittencourt, R. Bonefeld: *"Modeling machine tools for self- optimization of energy consumption"*, in: Proceedings of the 18th CIRP LCE Conference, Braunsch - Weig, 2011, pp. 253-257.

[10] A. Gludovátz, L. Bacsárdi: *"Production Related IT Solutions in the Operation of Factories"*, CINTI 2016, 17th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, Nov 17th-19th 2016, pp. 187-191.

[11] ISO 50001:2011(E). International standard, energy management systems – requirements with guidance for use. International Organization for Standardization; 2011.

[12] J.E. Seem: *"Using intelligent data analysis to detect abnormal energy consumption in buildings"*. Energy and Buildings 39 (1), 2007, pp. 52-58.

[13] C. Herrmann, T. Heinemann, S. Thiede: *"Synergies from process and energy-oriented process chain simulation e a case study from the aluminium die casting industry"*. In: Proceedings of the 18th CIRP International Conference on Life Cycle Engineering. Technische Universität Braunschweig, Springer-Verlag Berlin Heidelberg, Braunschweig, Germany, May 2nd-4th 2011a.

[14] K. Vikhorev, R. Greenough, N. Brown: *"An advanced energy management framework to promote energy awareness"*, Journal of Cleaner Production 43, 2013, pp. 103-112.

[15] C.-F. Lindberga, S. Tanb, J. Yanb, F. Starfelt: *"Key performance indicators improve industrial performance"*, Energy Procedia 75, 2015, pp. 1785-1790.

[16] F. Erdős: *"In-memory adatbázis-kezelő alkalmazások gazdaságossági kérdései"*, In: ENELKO 2015 XVI Nemzetközi Energetika-Elektrotechnika konferencia; SzámOkt 2015 XXV. Nemzetközi Számítástechnika és Oktatás Konferencia. Konferencia helye, ideje: Arad, Románia, 2015.10.08-2015.10.11. Arad: Erdélyi Magyar Műszaki Tudományos Társaság (EMT), pp. 203-207.

Business Information Visualization in Tangible Ways

[1]Ferenc Erdős – [2]Richárd Németh
[1]Széchenyi István University, Hungary; [2]T-Systems Hungary Ltd., Hungary
eMails: erdosf@sze.hu; nemeth2.richard@ext.t-systems.hu

Abstract

Data have a central role in corporate decision-making; the decision-making process could not be imagined without them. Business decision-makers want to rely on such analytical tools during the decision-making process that make the efficient use of data, information and knowledge achievable. By the growth of technological development and available performance, the development of the area of different visual presentations is becoming increasingly important. Within this area, 3D printing technology has made remarkable progress in the last decade considering the commercial successes and the availability of the technology as well. This study attempts to analyze the potential application opportunities of 3D printing in business information visualization.

Introduction

The traditional world view assumes that there are only two prime resources; material and energy. In fact, there are three of them – material, energy and knowledge. Knowledge comes from the huge amount of collected data to be processed – and from its essence: information. These days we have to deal with such an enormous amount of information that the processing of these exceeds human capabilities [8]. Just think about all those data which are related to stock exchange transactions, catalogs of public libraries or the stream of information required to manage a corporation that decision-makers have to understand. There are countless ways of storing, processing and displaying data, but the most obvious and most effective way of illustration for end-users is undoubtedly the visualized presentation.

Information visualization is a process by which the collected and processed data – placed into an appropriate environment – are presented by different methods so that they are understandable and transparent for the viewer. During the process information are visualized through forms and shapes. Technically spoken, information visualization is a process of mapping data to visuals. [15]

Figure 1. Information Visualization - A Spectacular Way of Presenting Information

The aim of data visualization is not only to display data in a spectacular and understandable way, but it is also equally important to make the significance of relations, relationships, changes and differences between these data easier for the user to understand and to make them transparent. In recent years, an assortment of techniques was introduced to visualizing complex features in data by relying on information abstracted from the data [4]. The primary goal is to promote effective processing using visual display devices. Researches prove that our brain can more easily absorb and process the visually depicted information, and it is easier to recall them later [13].

By now users require not only to see, but to perceive, to hold and to go around the tools of their work or the subject of their interest. If education is taken as the basis, several experiments prove that most of the students can learn more easily if they can tangibly experience information instead of merely reading or seeing them projected [16]. The same is true for the world of work; especially in professions where high precision and good spatial awareness are required.

The Information Visualization

While we generally consider information visualization (at least the term) as if it was a child of modern age, reality is much more subtle. If we start out from the definition of the term described in the previous chapter (namely that the purpose of data visualization is to represent collected data more expressively and efficiently), then we also have to list the geometric diagrams and astronomical maps drawn by Greek wise men in ancient times, the world map of Ptolemy, the maps drawn to warlords, or the anatomical drawings of Da Vinci here.

In the 14th century, Arabic numerals replacing Roman number representation made calculations much easier. The breakthrough in information visualization was represented by the new graphical methods that were spreading in the 18th century, primarily in the visual representation of geography, economics and medicine. The first more complex terrain maps, topography and graphs based on comparisons, as well as the representation by the help of geometric shapes appeared at that time [3].

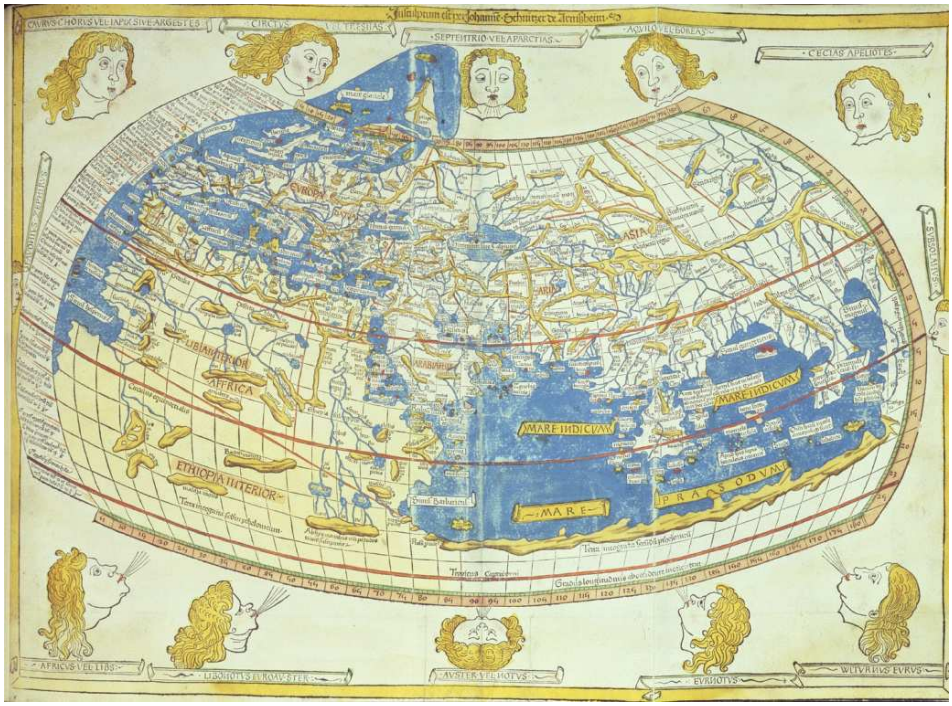


Figure 2. Ptolemy's World Map

The appearance of computers gave the next big boost; database management systems in data storage and data processing, various graphical applications and office suit packages in representation. Nowadays most infographics, line, circle and bar graphs are created with applications developed for this purpose.

The evolution of information technology and the many innovations opened up a new space in information visualization. In the area of technologies entitled for the extension of perception there has been a lot of progress in the last decades. The so-called “sonification”, designed for presenting, “amplifying” visual data and images [11], or vibration

(haptic) feedback (being used primarily in mobile phones) and the so-called "force feedback" (which is present in computer game controllers for many years) are also entitled to make the user's experience more complete. In this area real-time transformation, dynamic mapping and tangible/perceptible interactive communication would be a further step forward.

The Business Value

Information visualization has now become a standalone area, and one of the most popular areas of business intelligence. In the field of business intelligence, the purpose of data processing and visualization is mainly the support of decision-making. In doing so, it enables to the participants of the business process to display those information properly which are needed in decision making. Its main advantage is that it gives a better overview of the data in a flexible, configurable way; hereby it speeds up and makes more responsive the process of making business decisions.

With the modern data visualization solutions we can understand even more easily, more efficiently and faster what our customers really need, for example the time to enter to new markets can be reduced or the profit can be further increased. With the help of different BI solutions companies are able to define key performance indicators (KPIs) and monitoring them by visualization techniques that best reflect performance and operations. Through this, their processes and sales numbers are easier to understand, they can produce better forecasts, thus they are able to make better decisions, which ultimately have a positive impact on business outcomes. The representation of data (the form of storage) significantly determines the possibilities and effectiveness of visualization. The most common data representations are relational database tables and OLAP (On-Line Analytical Processing)-based data cubes.

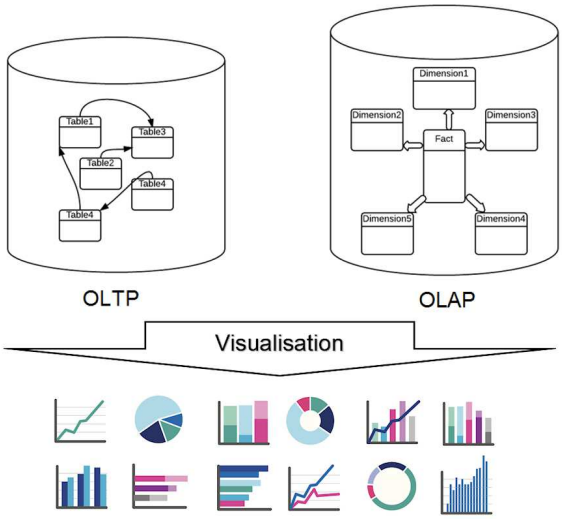


Figure 3. Information Visualization from Different Sources

The business value of information visualization is always related to corporate decisions. This means that the decision maker has a certain range of information and he wants to make these decisions better by obtaining new information. From this point of view, the adequate representation of relevant information can provide added value.

Acquiring all kinds of information (e.g introducing and maintaining an IT system) involves a certain amount of cost; and possessing information ensures a certain advantage for the decision maker. The cost of acquiring new information becomes questionable from an economic point of view when its input costs become higher than the value it provides, or the existing information is sufficient to make the right decision [14]. So, according to the rational model of economics, the acquisition of any information can be considered as uneconomic, which has no influence on decisions. In order to determine the exact value of information visualization, we need procedures which make the financial utility of given information determinable. Due to space limit this area is not the subject of this study.

Different Visualization Techniques

In contrast to the direct meaning of the words describing the concept, data visualization does not only visualize the data itself – we mean the visual representation of the data under the term, in order to facilitate the understanding of the context behind

the data, to bring the content and information they represent to the target as easily and quickly as possible. Consequently, the goal is not to primarily and exclusively present the concrete numerical values, but also the correlations, changes and differences. These can be illustrated by colors, sizes and shapes.

Nowadays the most common presentation methods are charts; for example this may be a line, pie or column chart (and much more – we only deal with the most widespread ones in this study). Line charts are usually used primarily to indicate growth or decline, bar charts are used for comparison of quantities and value sets and pie charts indicate ratios and shares. Of course, these are not strict rules, basically any form of presentation is suitable for displaying data, but we most often use them in the way described above [17].

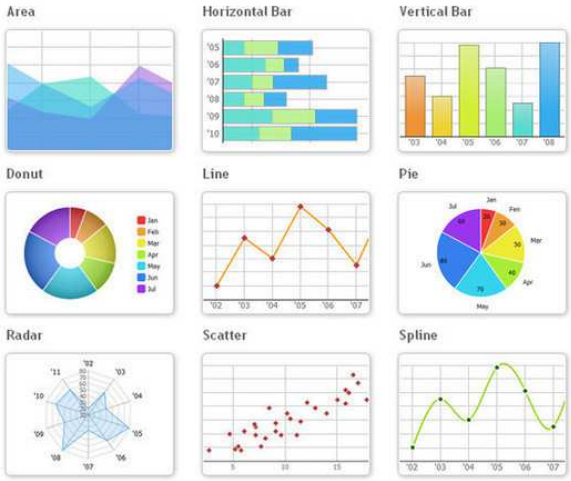


Figure 4. Most Common Simple Diagram Types

The above-mentioned methods are not suitable for representing complex data structures; in such cases we use so-called flow charts, link diagrams, text diagrams (e.g. word clouds or label graphs) or infographics if a few individuals are presented only.

Today, the IBCS (International Business Communication Standards) already provides practical recommendations for designing charts, reports and dashboards [10]. These standard recommendations include visual content conception, visual perception and semantic unification.

The display interface also basically determines the effectiveness of data visualization. At the dawn of literacy, the place of primitive figures drawn into stone slabs or terracotta was taken over by more complex signs in papyrus rolls; and later, the invention of the more foolproof and more durable parchment was an important step (170 B.C.). At the turn of the first century, paper has been discovered in China, and it came to Europe through the Arabs after hundreds of years of secrecy – and became the most effective platform for recording information for a very long time [7].

Although the manual (i.e. hand-drawn) technologies are now being replaced by graphs, infographics and others designed by target software on computer, we still often print these out. Unlike printed versions, using displays is a more flexible and cost-effective method, as we do not have to spend money (and time) for printing out updated data visualizations over and over again. The so-called e-paper technology has been in existence for nearly 30 years, and it has undergone significant improvements over the last decade. The purpose of this technology is to replace the printing on conventional paper by allowing almost an unlimited number of redrawing or reprinting (the e-paper requires a power source to change the content, but not for displaying) [9]. At this time, we can mainly see it as screens of electronic book readers and clock faces of smart watches. By now, newer tools of group communication, decision making, education and meetings have been developed, such as interactive boards, touch screens and tablets, as well as voting applications and software related to these. Colors also have a very important role in visualization. Certain colors have serious meanings that can vary by culture or subculture – for example if we mark something in red on a graph, some people are likely to find some problem behind the data, as red is often a sign of danger. In addition, there are fixed meanings for colors and signals in many fields of science (e.g. geological profiles or meteorological cyclones). The IBCS standards also propose concrete recommendations about the colors used in information visualization.

Figure 5. Multi-touch Interactive Tables

By the growth of technological development and available performance, the development of the area of different visual presentations is becoming increasingly important. Within this area, 3D technology has made remarkable progress in the last decade considering the commercial successes and the availability of the technology as well. The attention of the profession is also turning toward vision, especially toward 3D visualization; by way of example, systems based on virtual reality (VR), holographic displays or the novelty of the past decade, the additive manufacturing technology (AMT).

In modern business life, emphasis is increasingly shifting to spatial and physical information representation. To accomplish this, such software and hardware elements are required that enable the fast, efficient and inexpensive production of models needed for information visualization.

Technologies for 3D Models

Among production technologies aimed to create finished products, models, components and other solid objects, the so-called additive technologies are becoming more and more important. These procedures – in contrast to conventional, so-called subtractive methods (such as chipping, material separation, casting, melting or molding) are building up the target object from layer to layer. Additive technology can basically change the entire manufacturing industry. Nowadays mass production is still carried out by conventional methods; the rapid prototype manufacturing with 3D printing, while the rest of the process by injection molding or other alternative methods. However, for the personalized, small-scale model production we need, 3D printing is the ideal solution, mainly due to its flexibility, the ease of fabrication and the relatively low production costs. 3D printing does not “extract” the object from a given block by removing excess, but builds it up from its own material – and thus complex forms and shapes can be created, which would not be possible with other techniques.

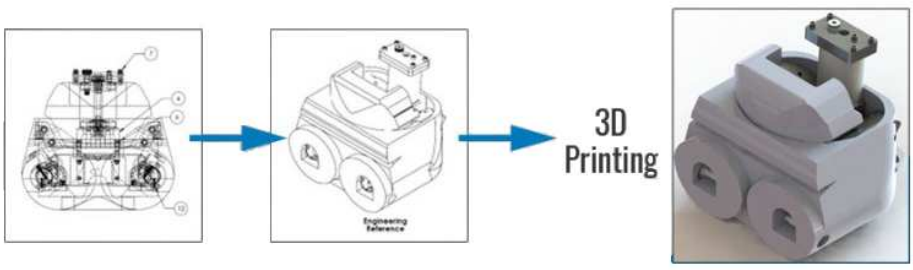


Figure 6. From Blueprint to Object

3D Printing in Model Production

3D printing is a type of manufacturing technology, which is basically about building up the final object in three dimensions, from layer to layer, using raw materials (these are usually so-called "filaments", but the raw material also can be metal, glass or plastic) melted at high temperature. The official term is Additive Manufacturing Technology – the word "additive" refers to the way of the printing process; the printer builds up the target object from layer to layer. "Manufacturing" suggests that this is a repeatable, plannable, automated and systematic line of actions [12]. During 3D printing – unlike other manufacturing processes – there are no wasted materials, and considering the method as a whole, it makes possible a faster, more economic and more complex production. On the other hand, AMT is not for mass manufacturing; it is more similar to the way how jewelers, sculptors or painters create their artworks. But while these artificers have to learn for a long period of time, the basics of 3D printing may be acquired quickly and relatively easily.

During printing, the machine creates the objects by the coordination of so-called "blueprints", which are design schemes of 3-dimensional models. These models can be created and modified via a design planning software known as CAD (Computer Aided Design), built up from polygons, which are digital mapping of 3-dimensional points in the space. The final objects are usually saved in STL (Stereo-Lithography) file format.

Figure 7. Conceptual Model of 3D Printing

STL is the standard industrial file type of printable 3D models, which contains the printable object's cross section in a structure built up from meshes (slices) [6]. Despite the fact that the technology has been available for more than 30 years, there were numerous major progresses made in the area in the last years.

As in the case of this process complexity does not influence time and material costs, it is capable of visually representing any type of data from the simple elements to the complicated, composite objects – without the shape restrictions of industrial fabrication machineries. There is no need for complex rework or assembly when the printing is ready, so this procedure is more efficient than any other ways of production [2]. The unbelievable flexibility of raw materials (which means that almost any type of materials can be used even with different surfaces) is an important factor, when data visualization is built up on physical touch.

The main disadvantage comes from the size; objects larger than the machine must be made in modular structure, from more parts. The combining process of colors and raw materials also requires a multistage manufacturing technique – however, many developments have been made in this area over the years and in the term of multi-material printers some significant innovations are being expected. These machines are able to print simultaneously from more raw materials, without having to chance them during the process [18].

All in all, there is a significant, but still unused potential in 3D printing (at least in this area), which makes the technology suitable for the quick and efficient presentation of business information transmission, in a way where not only spatial visual opportunities but the potential in physical touch and perception also can be harnessed to the fullest extent possible.

New Generation: 4D Printing

The raw materials we will use in the printing process have significant role in the model; the so-called 4D printing technology is already under development and practically workable; it provides a new dimension for the process of 3D printing, which is time – it means that the object is able to change its shape after manufacturing. The main thing in this innovation is the production of flexible, memorable, shape-changing materials. Those objects which are produced by this way can be moved or changed; what is more, they are able to transform, change their structure, or fix their own failures and damages [1].

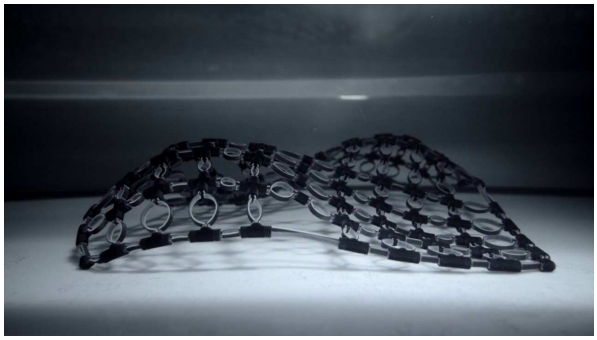


Figure 8. 4D Printed, Programmable Material

For us, the most useful feature of these objects printed by 4D technology is that the given structures can take up multiple stable forms; these materials are able to respond to changing circumstances (for example temperature, humidity, touch) by changing their own configuration – and this might be a huge step forward in the case of interactive communication mapping. Moreover, the shape changing is a particularly fast process [5]. Thanks to the hydrogels used, by the impact of the heater medium the change takes place within a few seconds, and the printed objects are able to "multiply" their original size.

For the tangible visualization of business information, this technology seems to be an excellent way, as the physical model can be updated quickly by changing the data source, which is the subject of visualization. Our conceptual model based on this statement is shown in Figure 9 below.

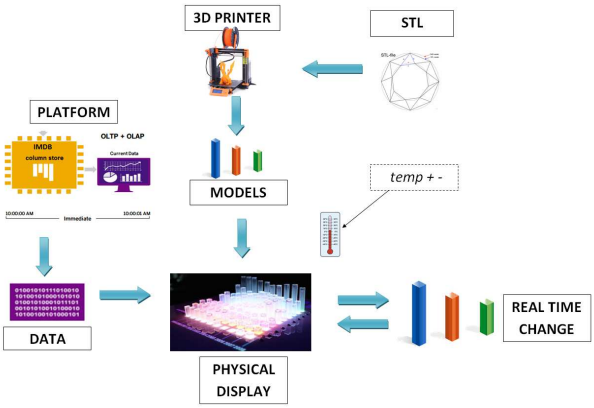


Figure 9. Business Information Visualization in Tangible Way

Based on this model, physically visualized business information can provide useful help to managers. Providing business-critical information in a physically tangible way gives a more pronounced impression and a more complex user experience than a conventional, classic dashboard-base solution.

Conclusions

With the development of visualization technologies, new opportunities for business information representation are emerging. In this study, we tried to focus on this novel field and to envision the potential inherent in the physical representation of business information. For the tangible visualization of business critical information, the additive manufacturing technology using shape-changing raw materials seems to be a suitable way, as the physical model can be updated quickly by changing the data source, which is the subject of visualization. With this visualization technique managers can not only to see, but to perceive, to hold and to go around the 3D-printed charts and infographics of relevant business information. The topic naturally requires further research. Our further aim is to develop a working prototype based on the model outlined above.

References

[1] Bakarich, S. E., Gorkin, R., Panhuis, M., Spinks, G. M. (2015) '4D Printing with Mechanically Robust, Thermally Actuating Hydrogels', Macromolecular Rapid Communications, 36 (12), pp. 1211–1217.

[2] Bradshaw, S., Bowyer, A., Haufe, P. (2010) The intellectual property implications of low-cost 3D printing. ScrpitEd, 7 (1), pp. 5–31.

[3] Chen, C., Härdle, W.K., Unwin, A. (2007) Handbook of Data Visualization. Springer, pp. 19-31.

[4] Chen, M., Ebert, D., Hagen, H., Laramee, R.S., van Liere, R., Kwan-Liu, M., Ribarsky, W., Scheuermann, G., Silver, D. (2009) 'Data, Information, and Knowledge in Visualization', IEEE Computer Graphics and Applications, vol. 29, no. 1, pp. 12-19.

[5] Choi, J., Kwon, O. C., Jo, W., Lee, H. J., Moon, W. (2015) '4D printing technology: A review, 3D Printing and Additive Manufacturing', 2 (4), pp. 159-167.

[6] Chua, C. K., Leong, K. F., Lim, C. S. (2003) Rapid Prototyping: Principles and Applications (2nd ed.). World Scientific Publishing Co, p. 237.

[7] Friendly, M. (2007) A brief history of data visualization. In C. Chen, Wolfgang Hardle and Antony Unwin, eds.,Handbook of Computational Statistics: Data Visualization, vol. III. (1), pp. 1–34.
[8] Harari, Y.N. (2015) Homo Deus: A Brief History of Tomorrow. Harper, p. 127.
[9] Heikenfeld, J. (2011) A critical review of the present and future prospects for electronic paper. J. Soc. Inf. Display. 19 (2): 129.
[10] Hichert, R., Faisst, J. (2014) Notation standards in business communication and their practical benefits. HICHERT®IBCS WHITE PAPER. Hichert Partner 2014. p 15.
[11] Hildebrandt, T., Rinderle-Ma, S. (2015) Server sounds and network noises, 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Győr, 2015, pp. 45-50.
[12] Lipson, H., Kurman, M. (2015) Fabricated – The New World of 3D Printing (2nd ed.). John Wiley & Sons, Inc., p. 65.
[13] Moulton, S. T., Türkay, S., Kosslyn, S. M. (2017) "Does a presentation's medium affect its message?" Plos One 12(10): e0186673.
[14] Müller, A. (1992) Informationsbeschaffung in Entscheidungssituationen. Verlag Wissenschaft und Praxis, Ludwigsburg – Berlin.
[15] Murray, S. (2013) Interactive Data Visualisation for the Web. O'Reilly Media Inc, p. 72.
[16] Schelly, C., Anzalone, G., Wijnen, B., Pearce, J. M. (2015) Open-source 3-D printing technologies for education: Bringing additive manufacturing to the classroom. Journal of Visual Languages & Computing, 28, pp. 226-230.
[17] Simkin, D., Hastie, R. (2012) 'An Information-Processing Analysis of Graph Perception', Journal of the American Statistical Association, 82:398, pp. 454-465.
[18] Sitthi-Amorn, P., Ramos, J.E., Wangy, Y., Kwan, J., Lan, J., Wang, W., Matusik, W. (2015) 'MultiFab: a machine vision asssisted platform for multi-material 3D printing', ACM Transactions on Graphics, Vol. 34 (4), pp. 1-11.

SEFBIS Board's Adopted and Performed Decisions of 2017

Decision No. 1/2017.

The GIKOF SIG Board discussed and accepted the Annual Workplan required by the Secretariat of the NJSzT. Decisions made on 17th of January, 2017:

- Acceptance of the Annual Report about activities in 2016 by the JvN CS Leadership
- Review and validate the Board and the SEFBIS SIG membership
- Actualize the SEFBIS Website: http://raffa6.wixsite.com/sefbis
- Cooperation in organizing the 14th OGIK Conference in Sopron (November 2017)
- Upload the SEFBIS Journal to the database of EBSCO for indexing
- Cooperation with other SIG groups of JvN CS
- Decision on the 15th OGIK/ISBIS Conf. 2018

Decision No. 2/2017

SEFBIS Board meeting at the General Assembly of JvN CS and made decisions as follows:

- Conference *date* and *venue*: 10-11. November 2017 Sopron University
- The International Programme Committee:
  Chair: RAFFAI, Mária (Széchenyi István University); *Co-chairs:* DOBAY, Péter (Pécs University of Sciences) and BACSÁRDI, LÁSZLÓ (Sopron University)
  Members: CHROUST, Gerhard (Johannes Kepler University, Austria); DOUČEK, Petr (Prague University of Economics, Czeh Republic), GÁBOR, András (Budapest Corvinus University); KŐ, Andrea (Budapest Corvinus University); KRUZSLICZ, Ferenc (Pécs University of Sciences); TJOA, A Min (Vienna University of Technology, Austria) and UCHIKI, Tetsuya (Saitama University, Japan)
- The *Organizing Committee* of the OGIK'2017 :
- *Chair:* BACSÁRDI, LÁSZLÓ (Sopron University)
- *Sponsors* of the Conference will be: JvN CS, Sopron University, Foundation Alexander, Guidance Ltd.

Decision No. 3/2017

SEFBIS SIG Board meeting held on 9th of November 2017 before the 14th ISBIS Conference

- The SEFBIS SIG Board expresses gratitude and thanks to organizers, reviewing board members and authors to successful work of the annual 14th OGIK/ISBIS Conference.
- Decision about the selection of the papers that will be published in the SEFBIS Journal No. 12 (in English) and about the date of sending the papers to the Committee by 28th February 2017 the latest. The papers, the contact with the authors, the whole reviewing process and the editing, printing works will be managed by Maria Raffai.

Decision No. 4/2017

The SEFBIS Board revised the members' activity and renewed the membership. The annual Activity Report'2017 and the Workplan of SEFBIS SIG for 2018 have been prepared. Decision was made about the circulation in December and sending to the Secretariat of the JvN CS by 20th December, the latest.

SEFBIS' Action-Plan for 2018

The Board of SEFBIS Special Interest Group

Chair: Mária RAFFAI
Members active in 2017: György BÖGEL, András GÁBOR, Zsolt KOSZTYÁN, Andrea KŐ, Zoltán VAJNA
Honorary members: Péter DOBAY, Gábor HOMONNAY

At the first meeting held in January 2018 the Board decided to perform the following tasks for the Year 2018:

Description	Estimated termin	Estimated venue	Participants, responsibility
Discuss, plan and accept both the action- and financial-plan for 2018	Until 15 January 2018		Members of the SEFBIS Board
Active participation at the meeting of the JvN CS professional communities, giving proposals for more effective cooperation with Secretariat and the other SIGs.	Februar 2018	Meeting Room of JvN CS	SEFBIS SIG Chair
Accepting papers to SEFBIS Journal No 12	28th February 2018 latest	NA	Editor in Chief
Reviewing process for paper having sent to SEFBIS Journal	31st May 2018	NA	Editor in Chief and reviewers
Call for publishing in the GIKOF and SEFBIS Journals Collecting the papers	continuous	NA, virtual	SEFBIS Board
Managing the work of reviewing and editing process of publishing the professional Journal(s).	continuous	NA, virtual	Approximately 12 reviewers
Update the GIKOF/SEFBIS website (content and design)	continuous	NA	SEFBIS Chair
OGIK/ISBIS'2018 conference: call for papers, applying for sponsorship, reviewing papers, organizing work	from April to November 2018	NA, virtual	12 colleagues, SEFBIS Board, IPC, experts
Cooperation both with leaders of business and university departments on BIS	continuous	face to face and virtual comm..	4-5 Board member Resp. SEFBIS pres.
Managing the editorial and printing work of the SEFBIS/GIKOF Journals; Uploading the new Journals to the international EBSCO Database	continuous	NA, virtual	Editor in Chief and Chair
Development and uploading database on experts, professionals in the field of BIS in Hungary	continuous	gathering information	2-3 Board members + activists, students
Preparing a competence map of professionals/teachers, Organizing on-line lectures for all BIS students in Hungary	continuous		2-3 Board members + professors
OGIK/ISBIS'2018 Conference	9-10 November 2018	Sopron University	60-80 participants + students, 40-45 papers
Evaluation of the yearly activity: results, report to JvN CS Presidency	18 December 2018 latest	NA, virtual	SEFBIS SIG Chair

International Conference on Research and Practical Issues of Enterprise Information Systems

Fudan University Shanghai, China // www.confenis.org

The 11th IFIP WG 8.9 Working Conference, the CONFENIS 2017 provided an international forum for Enterprise Information System's (EIS) researchers and practitioners from all over the world in order to gather, present and discuss their latest research results and findings. The conference aimed to facilitate the exchange of ideas and developments in all aspects of EIS. The theme of CONFENIS 2017 was the *Industrial Internet of Things* and *Made in China 2025*. The Workshop of Smart Electronics and Systems for Industrial IoT was held jointly with the conference. The world's public authorities, industries, researchers and academia were cordially invited to participate in this event in October, one of the best months to visit Shanghai, which provides the best climate.

Topics of interest of the CONFENIS 2017 included, but were not limited to:	Topics of Special Theme on Smart Electronics and Systems for Industrial IoT:
– EIS Concepts, Theory and Methods – Business Process Management – Enterprise Architecture and Engineering – IoT and Emerging Paradigm – Cyber-Physical Systems (CPS) and EIS – EIS for Industry 4.0 – Industrial Digitalization and Big Data Analytics – EIS Management and Case Study	– Smart Electronics and Devices – Smart Systems Integration – Chip-Cloud Integration – Fog and Embedded Intelligences – Smart CPS – AI for Industrial Informatics – Industrial Information Integration

Publication

The best papers of CONFENIS 2017 has been published in the Springer Lecture Notes on Business Information Processing (LNBIP) Series. Other publication outlets are now also available in the Journals listed below:

- • Journal of Industrial Information Integration (Elsevier BV, Scopus and Inspec)
- • Journal of Management Analytics (Taylor & Francis, edited by Shanghai Jiao Tong University)
- • Journal of Industrial Integration and Management (World Scientific, Singapore)
- • Nanotechnologies in Construction (ESCI)

The conference was hosted by the Fudan University in Crowne Plaza in Fudan district of Shanghai, China. The chair and members of Organizing Committee did their best also in advance in order to manage an event that gives not only opportunity for the professional from all over the world to present and discuss the newest results and technologies of ICT, but it was also a good occasion to meet and chat freely during the evening programs.

The opening ceremony of the conference was followed by interesting keynote speeches on October 19, in the huge Ball Room of Crowne Plaza Hotel. The room was full with participants who were very much interested in the different and challenging topics of the speeches. Let us see the speakers and the title of the presentations:

- Maria RAFFAI (IFIP Councilor, Vice Chair of WG 8.9, professor at the Széchenyi University, Hungary): China's Commitment on Enterprise Information Systems
- Jianguo WANG (General Manager of Bosch Software Innovation, China): Bosch IoT Suite – The Foundation of Industry 4.0
- Mikhail Yurevich KATAEV (Professor at Tomsk State University of Control Systems and Radioelectronics, Russia): The Business Process Approach in Management of Enterprises
- Nina Maarit NOVAK (University of Vienna, Austria): The Economic Value of an Emergency Call System
- Antonin PAVLIČEK – Petr DUČEK (University of Economics, Prague, Czech Republic): Big Data Analitycs – Geolocation from the Perspective of Mobile Network Operator
- Rudy LAUWEREINS (Vice president at IMEC, professor at the Katholike Universiteit Leuven, Belgium): Mass-Manufacturable Advanced Electronics to Address Grand Challenges in Health Care
- Attila ALVANDPOUR (professor Lindköping University, Sweden): Microwatt Piezoelectric Energy Harvesting and Smart Sensors for Internet of Everything

The three days conference continued in two parallel tracks with four sessions in each. The authors presented their works and discussed them with the participants. We give you a short overview about the most interesting issues:

(1) the main topics of *the solutions and the way of usage of intelligent electronic systems* track:

- mobile code approach of interplaying in IoT,
- high sensible wireless communication technology for underwater Internet systems,
- effectiveness and efficiency of vehicle instance retrieval,
- TV panels' monitoring for quality assurance,
- security framework for fog networks
- optimization of energy consumption
- ensuring fault tolerance and efficiency in Swarm-based monitoring system
- survey and analysis on Swarm of unmanned aerial vehicles

(2) the topics of *Different Views of Enterprise Information Systems* track:

- penetration of Industry 4.0 principles into ERP products – a Central European study
- method of domain specific code generation based on knowledge graph
- modeling of service time in public organizations based on business processes
- automated trading systems, machine learning methods, hardware implementation
- data analysis in the healthcare of a smart city
- IoT platform for real-time multichannel ECG monitoring with neural network
- checking the correctness of *What-if* scenarios
- a survey on using enterprise architecture: principles, methods, models
- aquatic product traceability platform for EIS management
- image database management architecture: logical structure and indexing methods
- analysis method towards product quality management

- an IoT – Big Data machine learning technique for forecasting water requirements in irrigation field
- Internet of Things or Surveillance of Things
- An architecture for integration of smart city applications with Legos
- Systematic analysis of future competences affected by Industry 4.0
- Genomics Cloud design and implementation
- Big Data analytics using SQL
- modeling the popularity of online topics based on survival analysis

The conference was an event of the IFIP WG 8.9 Enterprise Information Working Group, so the chair A Min TJOA and the vice-chair, Li Da XU took the responsibility as honorary chairs of the Organizing Committee for the high professional value of the conference. The professors at Fudan University, Zhou ZOU and Lirong ZHENG directed the organizing team in managing all the important tasks that had to be done during several months in advance the conference. The International Program Committee chairing by Maria RAFFAI reviewed the papers, organized and monitored the professional program.

We can conclude that the CONFENIS 2017 conference that was hosted already second time by Chinese universities was really a successful conference that is worth to organize every year. It gathers all the professionals interested and active in Business Information Systems and Enterprise IS from every continent, and gives opportunity for the researchers, developers to present and discuss their results and the future of ICT.

As the chair and vice chairs of the IFIP Working Group on Enterprise Information Systems we express our thanks and acknowledgment to all the colleagues at the Fudan University for giving us the opportunity to spend 3 wonderful days not only at the conference but also with discovering and enjoying the wonderful city of Shanghai! See you next year in Poznan at the CONFENIS'2018 conference!

Li Da XU	*Maria RAFFAI*	*A Min TJOA*
Vice-Chair of WG 8.9	Vice-Chair of WG 8.9	Chair of WG 8.9

Conferences Organized Worldwide and Relevant to the SEFBIS Community

Event	Date/Location	Organizers' Contact
ISBIS–OGIK 2019 16th International Symposium on Business Information Systems	8-9 or 15-16/11/ 2019 ……., HU	NJSZT GIKOF SIG raffa6.wix.com/sefbis#!conferences
CONFENIS 2019 – International Conference on Research and Practical Issues of Enterprise Information Systems	15-16/10/ 2019 Prague, CZ	IFIP WG 8.9 petr.doucek@vse.cz
I3E 2019 eBusiness, eServices, eSociety	18-20/09 2019 Trondheim, NO	IFIP TC6, WG 6.11 ilpappas@ntnu.no
INTERACT 2019 – International Conference on Human-Computer Interaction	02-06/09 2019 Paphos, CY	IFIP TC13 pzaphiri@gmail.com
MIM 2019 – 9th IFAC Conference on Manufacturing Modelling, Management and Control MIM	28-30/08 2019 Berlin, DE	IFIP TC5, WG 5.7 divanov@hwr-berlin.de
DisCoTe – International Federated Conference on Distributed Computing Techniques	18-21/06 2019 Kongens Lyngby,	IFIP TC6, WG6.1 albl@dtu.dk
WiOpt 2019 – The 17 Internation Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks	27-31/05 2019 Avignon, FR	IEEE Control Systems Society elazouzi@univ-avignon.fr
15th AIAI 2019 Artificial Intelligence Applications and Innovations	24-26/05 2019 Kos island, GR	IFIP TC2, WG2.2, WG12 liliadis@civil.duth.gr
FSEN 2019 Fundamentals of Software Engineering	01-03/05 2019 Tehran, IR	azad@ipm.ir
ISBIS–OGIK 2018 15th International Symposium on Business Information Systems	9-10/11/ 2018 Sopron, HU	NJSZT GIKOF SIG raffa6.wix.com/sefbis#!conferences
I3E 2018 – e-Business, e-Services, and e-Society	29-31/10 2018 Kuwait City, KW	IFIP TC 6 AlsharhanS@gust.edu.kw
CONFENIS 2018 – International Conference on Research and Practical Issues of Enterprise Information Systems	18-19/09/ 2018 Poznan, PL	IFIP WG 8.9
Internet of Things Conference in the frame of IFIP WCC 2018	18-20/09 2018 Poznan, PL	IFIP Joint Event strous@iae.nl
WCC 2018 IFIP World Computer Congress	17-21/09/2018 Poznan, PL	IFIP
The Dewald Roode **Workshop** on Information Systems Security Research	14-15/06 2018 Cape Town, ZA	IFIP WG8.11, WG 8.13 jacques.ophoff@uct.ac.za
INCOM2018 – 16th IFAC Symposium on Information Control Problems in Manufacturing	11-13/06 2018 Bergamo, IT	IFIP Supported Event marco.macchi@polimi.it
OSS 2018 International Conference on Open Source Systems	08-10/06 2018 Athens, GR	IFIP Event varlamis@hua.gr

Report on ISBIS/OGIK 2017 10-11th of November, Sopron

International Forum of Scientific and Educational Forum on Business Information Systems

MÁRIA RAFFAI
chair of International Program Committee; *eMail:* raffai@sze.hu

Reviewing the applications for the right of organizing the 14th ISBIS/OGIK Conference, the Board decided to give the opportunity to the Sopron University. The colleagues at the university were very much committed in planning, managing and performing the OGIK conference in Sopron. They found new solutions, new forms of presentations in order to make this event more effective and attractive. László Bacsárdi, the chair of Organizing Committee and his eager colleagues such as Gergely Bencsik, Attila Gludovátz, Péter Kiss, László Koloszár, Zoltán Pődör and Mónika Tóth have been constantly ready to arrange everything in connection with the conference. Nevertheless, not only the Organizing Committee worked hard during 8 month before the conference, but also the members of the International Program Committee, who reviewed the papers, and managed the professional program. Anyway, the hard work finally fructified, the authors have sent interesting and valuable papers, and the presentations arouse a high interest on the different topics.

The 14th OGIK Conference was organized in Sopron on 10-11th of November 2017 in Ligneum Conference Centre of the Sopron University. The conference motto had been selected as 'ICT in the daily work" inspired the authors to submit papers dealing with results and experiences in the field of ICT innovation. Members of the Program Committee controlled altogether 43 subscriptions with a double-blind review process. They accepted 15 papers for conference presentation and 19 for poster presentation. The lectures were organized in a plenary and 2 main sessions: Applications in BIS, Information Management, a Round table discussion and a Poster presentation.

After the official opening speeches the "*plenary session"* was introduced by the representative of NetAkademia. PÉTER LITKEY talked about the importance of the security in the era of Industry 4.0. The colleagues from the Opel Szentgotthard (CSANAKI-EDELÉNYI-KÁROLYI NAGY) presented how the company Opel utilizes the business intelligence. As during all the previous conferences, the problems and solutions of ICT education and training played important role among the presentations. ZOLTÁN GÖRÖG was talking about the reasons, why it is necessary to teach ERP, and MARTIN KAINDHOFER detailed and emphasized the importance of the common EDLRIS Austrian-Hungarian project for developing European Driving Licence curriculum in the field of artificial intelligence and robotics.

The afternoon and the Saturday program offered different sections to the audience:

- "*Section #1: Business Information Applications*" presented lectures on applications which support the performing business processes and help the users to perform the every day work faster, more effective and precise or even automatically. The main topics focused on testing and validation of ICT applications, on technologies for creating IoT infrastructure, on agent-base simulation software for application management and on the newest way for using ERPs.
- *"Section #2: Information Management"* has ever been a hot topic for our forums. We could hear interesting presentations about solutions, ICT support in eServices, about the users' requirements and behavior, about the maturity of information security and about other interesting topics. The presentations were followed with questions from the audience and even discussions.
- The *"Poster presentations"* focused on different topics the authors had been dealing with, such as IoT tools, mobile and Internet communication, developing methodologies, Industry 4.0, results on text-mining, data management and so on. All authors got 1 minute time for presenting their topic and calling the attention of the audience, who could later give questions and initiate debates to the authors in front of the poster-table. This kind of "poster-style" was really interesting and the participants found it very effective.
- *"Round Table Discussion"* was searching for answers of the role of industrial partners in BIS education and training. The colleagues from Feki Webstudio, Opel Szentgotthard, NetAkademia, GySEV ZRT and MVMI ZRT. emphasized that the industry needs professionals who are prepared not only for the newest technologies and solutions of ICT but also for understanding the business processes in order to develop and tailor well usable applications.

As conclusion, the situation needs cooperation of all partners who play important role in developing and using ICT products. The speakers and participants reported problems which should be solved and/or new technologies, solutions that serves the innovation. Consequently this needs urgent action, and the conferences like the OGIK and the common thinking helps us in reaching our goals! Participants closed the Conference with acknowledgement for the smooth management and hospitality of Sopron University and for all the colleagues who played active role in organizing, managing and performing the OGIK'2017 conference!

Mária Raffai
Chair of the IPC

International Conference on Research and Practical Issues of Enterprise Information Systems

CONFENIS 2018 provides an international forum for Enterprise Information System (EIS) researchers and practitioners from all over the world to come together, present and discuss their latest research findings and ideas. The conference is specifically aiming at facilitating the exchange of ideas and advances in all aspects and developments of EIS. The Conference invites EIS-experts who are interested in presenting and disseminating their work at an international forum. The proceedings of the conference will be published by Springer LNBIP as part of the WCC-2018 Conference Proceedings.

The Venue

CONFENIS 2018 will be held in the frame of IFIPs World Computer Congress-2018 (WCC 2018) at the Campus of Poznan University of Technology, Lecture and Conference Center, Poland from 17 – 21 of September 2018. On a two-year basis the International Federation for Information Processing (IFIP) brings together experts from all over the world, representing the commercial, industrial and scientific sector, to showcase and discover innovative ideas. Our Conference offers the unique possibility for researchers and participants to both attend and submit contributions to several conferences, co-located under the umbrella of the WCC'2018.

The CONFENIS 2018 is organised with the intent:

- to provide an international academic platform for scholars to exchange ideas and latest research results in the field of Enterprise Information Systems.
- to foster long-term relationships among and with researchers and leading organizations worldwide.
- to connect talented delegates from all over the world with leaders in academia, industry and government.
- to further investigate the huge potential of novel EIS developments.

Publication: The Proceedings will be published in the Springer Lecture Notes in Business Information Processing (LNBIP) Series.

For further information on the WCC'2018 please see http://wcc2018.org and on the CONFENIS'2018: http://webcampus.ifs.tuwien.ac.at/confenis2018/

ISBIS / OGIK Conference'2018

9-10 of November 2018; SOPRON

15th Conference on Business Information Systems

http://www.ogik2018.hu

The SEFBIS (GIKOF) Special Interest Group of the John von Neumann Computer Society organizes its conference on Business Information Systems already 15th time. This event is followed with great interest both of the professionals and the users. The conference is a great opportunity for the domestic and foreign experts, developers, students and users to represent and discuss their results, to change their minds and get together face to face.

The planned topics those are not limited to:

- New technologies supporting business innovation and competitiveness
- Business process modeling; model driven architectures
- System engineering concepts, methods and tools, modeling languages
- Knowledge based systems, fields and efficiency of applications
- Enterprise wide integration → integration of data, database and applications
- Big Data management: storing and processing, data mining; analysis
- Using open source solutions in the business, effectiveness of the applications
- Business value of the ICT, studying efficiency: analysis' methods and results
- Industrial solutions: challenges of Indutry 4.0 and IoT – software and architectures
- Multidisciplinarity: ICT applications medicine, self-driving cars, utilization of visualization etc.
- Education: CSc, MSc courses on business information systems, trainings, labour market needs vs teaching curriculums, changing competences, change management in education, teaching methods and tools, mobility

The authors and speakers are asked to send an extended abstract in max 4,000 characters, it should be sent through Easychair CMS in English or in Hungarian. The official language of the conference both Hungarian and English. The authors have to present their results in the language of the abstract. It is the IPC's responsibility to let the abstracts review and decide about the acceptance. The IPC offers different types of presentations:

• live presentation in 20 minutes
• poster demonstration

Important dates:

- 25th September 2018. abstracts uploading
- 1st October 2018: abstracts acceptance
- 9-10th November 2018: conference in Sopron University

The International Program Committee and Conference Organizers

- Chair of the IPC: Maria Raffai (Széchenyi István University, Győr)
- Vice Chairs: Péter Dobay (Pécs University of Sciences) and László Bacsárdi (Sopron University)
- Chair of the Organizers Committee: László Bacsárdi

Further details and on-line registration: www.ogik2018.hu (in HU); http://raffa6.wixsite.com/sefbis/ogik-2018 (in EN)

In the hope that this conference will also have as great interest as all the others in the previous years, on behalf of International Committee I invite all the professionals who are involved in Business Information Systems let them work either in business, in software development or in education, to participate at ISBIS'2018!


Scientific and Educational Forum on Business Information Systems
S·E·F·B·I·S