The 4th IEEE International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems
20-21 September, 2018, Lviv, Ukraine

Combined Models for Forecasting the Air Pollution Level in Infocommunication Systems for the Environment State Monitoring

Alexander Kuchansky[1], Andrii Biloshchytskyi[1], Yurii Andrashko[2], Vladimir Vatskel[3], Svitlana Biloshchytska[3], Olena Danchenko[4] and Igor Vatskel[1]
[1] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
[2] Uzhhorod National University, Uzhhorod, Ukraine
[3] Kyiv National University of Construction and Architecture, Kyiv, Ukraine
[4] KROK University, Kyiv, Ukraine
*Corresponding Author's email: yurii.andrashko@uzhnu.edu.ua

***Abstract* — The combined models of selective and hybrid types with indexation of time series for forecasting the level of air pollution in infocommunication systems for monitoring the state of the environment are developed. Indexing in these models takes place on the basis of the nearest neighbor's method with selected metric distances. Described models allow to achieve a higher precision of short-term and medium-term forecasting compared to the models included in the base set of these combined models. Models and appropriate methods can be used in the development of infocommunication systems for monitoring the state of the environment and hardware-software complexes of general environmental monitoring.**

Keywords — combined forecasting models; air pollution; environmental monitoring.

I. INTRODUCTION

Air pollutants can cause a wide variety of problems, including reduced visibility, unpleasant odors, damage to cultivated plants, and adverse effects on human and animal health (cardiovascular, nervous diseases, damage to the respiratory system, eyes, kidneys and other organs). They can influence not only the air, but also can pollute water and food. The overall quality of our environment in some cases can seriously influence the survival of humans, animals and plants. The effect of the influence the environment depends on the nature and extent of air pollution sources, the place and height of emissions, the emergence of the chemical transformation, on the meteorological factors too. Control and forecasting of air pollution level are necessary in order to protect the environment and human health.

The research task is to develop the combined models of selective and hybrid types with time series indexation for calculating the short-term and medium-term forecast of air pollution levels. It is assumed that these models can help to effectively detect and predict pollutants the concentration of which increases and can be harmful. This will allow to realize the operational management of various aspects of environmental safety in one or another territory (in city, industrial zone, etc.). The state of the environment is dynamically changing, therefore, the operation of this object can take place in some infocommunication system.

Infocommunication system for monitoring the state of the environment is a combination of:

- hardware-software complex for environmental monitoring. The hardware part of this complex is an ecological station. It is intended for continuous monitoring of various ecological indicators of the environment, pre-processing of information, its preservation and transmission to the program part with the help of certain communication channels. The software part is developed with the using of micro-service architecture and is intended for collecting, processing and storage of information from the network of ecological stations, its analysis and display to end users;
- methods of incoming information processing and forecasting models of air pollution levels;
- sources of information (information on air pollution by various pollutants, GPS coordinates from the place of collection data on the state of environment);
- consumers of information (organizations that provide environmental safety).

With the growth of industrial capacities and increase of countries development level, the task of air pollution forecasting becomes more relevant. In [1], the application of neural network models to the prediction of the concentration of some pollutants in the air of London has been investigated. In [2], the authors also applied the neural network to this task with the possibility of detecting periodic components in changing the concentration of pollutants. In [3], the method of reference vectors for short-term forecasting of air pollution level in Macao is considered.

The article [4] describes the methodology for obtaining key parameters from the weather forecast and the creation of a physical means for forecasting the air pollution status of NO_2. The authors found the existence of a significant periodic dependence of pollution indicators on the season, day of the week and time of day. In [5], the authors describe the system of notification of the increase in the level of pollution in an urbanized city. In [6] describes the difficulties that arise when development a real-time Valencia air monitoring system.

In this research the development of combined models of selective and hybrid types with pre-indexed of time series is considered. In [7] features of similarity detection in time series were considered. In [8, 9] the using of the nearest neighbor's method for forecasting of time series, in particular financial, were considered. In [10] the forecasting method, which uses the comparison of time series different fragments with the sample, is considered. In [11] forecasting method based on selective comparison with the sample is considered, possible criteria for selection of forecasting models are described. Forecasting method of signs increment, which uses the principles of constructing combined models, is described in [12].The method of the scientific directions potential forecasting in infocommunication systems of an assessment of the research activity results is described in [13]. Methods of constructing fuzzy expert systems are described in [14]. They can be used for effective forecasting, when choosing model parameters or at the stage of making decisions about trends in changing the time series values [15]. Application of adaptive methods of forecasting is described in more detail in [16].

II. THE PRIMARY RESEARCH MATERIALS

The level of air pollution by specific pollutants (carbon monoxide, bicarbonates, sulfur dioxide, nitrogen dioxide, lead, mercury, etc.) fixed at certain moments of time can be represented as a finite discrete time series without spaces. It is a sequence of real numbers that denote the concentration of the pollutant $r_i \in R$, $i = \overline{1,n}$

$$R = \left(r_1, r_2, \ldots, r_n\right) \tag{1}$$

In this case, the time series represents the levels of air pollution by one of the pollutants (carbon monoxide, carbon dioxide, dioxin, nitrogen dioxide, lead, mercury, etc.), which is fixed by the hardware-software complex of the infocommunication system for monitoring the state of the environment. Values can be fixed at intervals per month, week, day, hour, etc. The task is to calculate the level of air pollution by one of the pollutants with the horizon $\theta > 1$, i.e. for each next time point $n+1, n+2, \ldots, n+\theta$. In other words, it is necessary to develop a model for calculating estimates of the level of pollution with the period $\tau = \overline{1,\theta}$, i.e. to calculate the value of the forecast time series

$$R^* = \left(\bar{r}_{n+1}, \bar{r}_{n+2}, \ldots, \bar{r}_{n+\theta}\right), \tag{2}$$

where θ is fixed before the calculating of forecast.

Let m is the size of retrospective sample, that is, the dimension of input time series area, which directly follows the point at which the forecast is calculated (point r_n) and which is involved in the calculation of forecast estimates, $m \le n$. Functional dependence, on the basis of which the estimates of the forecast (2) are determined, is called the forecasting model. Moreover $\bar{r}_{n+\tau}$ is the forecast estimate, which is calculated at the point r_n for τ points forward or with a period τ, $\tau = \overline{1,\theta}$. If formally mark such a model through f, then the forecast, which is calculated at the point r_n on one point ahead or with period 1, can be defined in this way:

$$\bar{r}_{n+1} = f\left(r_{n-m+1}, r_{n-m}, \ldots, r_n\right) \tag{3}$$

Trend models, adaptive polynomial models of smoothing, etc. can be used to predict such time series. To calculate the forecast, let's consider the approach of combined models constructing, taking into account similarities in the dynamics of the input time series, as well as other time series of air pollution level. Consider two types of combined models: selective and hybrid. The forecast in a selective combined model is realized on the basis of a single model, which is elected by selective choose from the base set of models. The selective election procedure is implemented on the basis of a selection criterion. The selection criterion usually is an estimate of the forecasting error of this model at the forecast point. In this study, we will carry out a selective election separately for each of the values $\tau = \overline{1,\theta}$. This is due to the fact that some models are more accurate for short-term forecasting, others for medium-term.

Let's set the basic set of models for predicting the air pollution level $F = \left(f_1, f_2, \ldots, f_v\right)$. For each model f_k, $k = \overline{1,v}$ let's calculate the performance criteria for $\tau = \overline{1,\theta}$ according to one of the formulas:

$$\Phi_0^k(\tau) = \sum_{i=0}^{m} b_i^k \left| r_{n-m+i} - \bar{r}_{n-m+i}^{(\tau+1),k} \right|, \tag{4}$$

$$\Phi_1^k(\tau) = \sum_{i=0}^{m} b_i^k \frac{\left| r_{n-m+i} - \bar{r}_{n-m+i}^{(\tau+1),k} \right|}{r_{n-m+i}} \cdot 100 \tag{5}$$

where $\bar{r}_h^{(\tau+1),k}$ are predicted values of the time series, which are projected for τ points ahead by k model f_k, $k=\overline{1,v}$, $h=\overline{n-m,n}$, a b_i^k, $i=\overline{0,m}$ are normalized weights, $b_0+b_1+\ldots+b_m=1$.

For forecasting at the point r_n for $\tau=\overline{1,\theta}$ the model is selected from the set F, for a fixed m, $m<n$ and $j\in\{0,1\}$, for which the following condition is fulfilled:

$$f^{*,\tau}=\arg\min_{k=\overline{1,v}}\Phi_j^k(\tau) \quad (6)$$

where $f^{*,\tau}$ are models that are selected for forecasting for τ points ahead or with a period τ, $k=\overline{1,v}$, $v=card(F)$. The corresponding estimates of forecasts for these models will be marked as follows: $\bar{r}_{n+1}^{*},\bar{r}_{n+2}^{*},\ldots,\bar{r}_{n+\theta}^{*}$.

We will index the time series R. For this we consider a section of a series that precedes the forecast and includes the point at which the forecast is calculated:

$$r'=\left(r_{n-\mu},r_{n-\mu+1},\ldots,r_n\right) \quad (7)$$

where μ dimension of this plot. The value μ is determined experimentally or recorded on the basis of an expert survey. If μ was set too high, then the obsolete time series values will affect the result of the indexation and accordingly the forecast. If μ is too small, then some information about the behavior of the time series is lost.

Consider two options for indexing:

1. Finding such a plot of time series R with dimension μ, which is similar to a plot r' based on a certain degree of proximity, for example, the distance from Euclid (8), the measure of Minkowski (9):

$$, \quad (8)$$

$$w_1(R,r')=\left(\sum_{h=0}^{n-2\mu}\sum_{j=1}^{\mu}\left|r_{h+j}-r_{n-\mu+j}\right|^p\right)^{\frac{1}{p}}. \quad (9)$$

2. Putting into consideration another time series $Z=(z_1,z_2,\ldots,z_n)$, which also shows the level of air pollution and finding, on the basis of a certain degree of proximity to such a site of a given time series of dimension μ, which is similar to a plot r'. In this case, formulas (8), (9) will be:

$$w_0(Z,r')=\sqrt{\sum_{h=0}^{n-\mu}\sum_{j=1}^{\mu}\left(z_{h+j}-r_{n-\mu+j}\right)^2}, \quad (10)$$

$$w_1(Z,r')=\left(\sum_{h=0}^{n-\mu}\sum_{j=1}^{\mu}\left|z_{h+j}-r_{n-\mu+j}\right|^p\right)^{\frac{1}{p}}. \quad (11)$$

Let the area of dimension μ be determined on the basis of the first or second approach, which is similar to r'. Let's denote it through:

$$r''=\left(r_{n-\mu-g},r_{n-\mu-g-1},\ldots,r_{n-g}\right) \quad (12)$$

where $g>\mu$, $g+\mu\le n+1$.

Formally, this means that among other parts of the dimension μ an input time series or a set of time series there are no such ones for which the degree of proximity would be less than the degree of proximity between the sections r' and r''.

The forecast of the time series of the air pollution level $R=(r_1,r_2,\ldots,r_n)$, which is calculated at the point r_n with the horizon $\tau=\overline{1,\theta}$ on the basis of a combined model of selective type is determined by the formula:

$$\bar{r}_{n+\tau}=\rho\cdot r_{n-g+\tau}+(1-\rho)\cdot\bar{r}_{n+\tau}^{*} \quad (13)$$

where $\rho\in[0,1]$ is a parameter that determines which of the predictions (based on selection or based on the similarity of the time series plots) has a greater weight in determining the prediction result. Details of this model are described in [11].

When constructing a hybrid-type hybrid model, it is required for each value $\tau=\overline{1,\theta}$ to construct such sets of models, we will denote them through F_τ, which include only those models that are potentially the most accurate at the current time zone and $F_\tau\subset F$ at $\tau=\overline{1,\theta}$. Such a selection is used to ensure that less accurate patterns from the base set do not affect the prediction result [12].

We introduce the threshold value γ, which is determined by the predictor based on the results of the experimental part of the time series prediction. Then, by formula (4) or (5), we find the values of the errors and determine the sets F_τ by the rule:

$$F_\tau=\left\{f^k\,\middle|\,\Phi_j^k(\tau)\le\gamma,k=\overline{1,v},j\in\{0,1\}\right\} \quad (14)$$

that is, only those models for which the error value does not exceed the threshold value are selected γ.

Let's mark forecasts for models that are included in the sets $F_1, F_2, \ldots, F_\theta$ respectively through

$$\bar{r}_{n+1}^{(1),c_\tau}, \bar{r}_{n+1}^{(2),c_\tau}, \ldots, \bar{r}_{n+1}^{(\theta),c_\tau} \quad (15)$$

where $c_\tau = \overline{1, card(F_\tau)}$. Then the forecast for the combined model of the hybrid type taking into account such areas of the input time series R is calculated by the formula:

$$\bar{r}_{n+\tau} = \rho \cdot r_{n-g+\tau} + \\ + (1-\rho) \cdot \left(\sum_{i=1}^{c_\tau} \varphi_i \right)^{-1} \cdot \sum_{j=1}^{c_\tau} \left(\varphi_j \cdot \bar{r}_{n+\tau}^{(\tau),c_\tau} \right) \quad (16)$$

where $\rho \in [0,1]$ is a parameter that defines the weight of forecasts, and ϕ_i is a weight coefficients, $i = \overline{1, c_\tau}$, $\tau = \overline{1, \theta}$.

The algorithm based on models of hybrid and selective types taking into account the similarities of time series in the infocommunication system for monitoring the state of the environment consists of the following steps:

1. Loading time series R of level of air pollution by one of the pollutants.

2. Development of a basic set of models. For example, the base set can include adaptive Holt models of various orders, models of fluid averages, etc.

3. Selection the value of the prediction horizon θ and parameters values for forecasting models. Realization of forecasting models from the base set.

4. For each $\tau = \overline{1, \theta}$ on the basis of the base set formed a subsets of the most precision models in the current section of the series, that is, sets according to (14).

5. Time series indexation for identifying similarities. As the basis, the area of the input time series of dimension μ is selected, which directly follows the point at which the forecast is executed. Indexing is performed by the nearest neighbor method. As one measure of similarity, one of the metric distances is chosen (see formulas (8) – (11)). At the exit we get some section of the input or other time series, which is similar to the area that precedes the forecast.

Figure 1. The forecasting results in the Infocommunication system for monitoring the state of the environment (Inspector Meteo)

6. Calculation of the forecast with the horizon θ on the basis of a combined adaptive model of selective type taking into account indexation by (13) and a hybrid-type hybrid model with indexation by (16). The smoothing parameter ρ is selected experimentally. At the output we get a forecast time series of air pollution levels R^*.

III. RESULSTS OF RESEARCH

The result of research is the formalization of combined models of selective and hybrid types with indexation of time series for forecast the level of air pollution in infocommunication systems of environmental monitoring (Inspector Meteo).It was found that in the case of use of a selective prediction model with time series indexing, it is possible to obtain a higher prediction efficiency compared to models included in the general set of values $\tau \geq 3$. Using the same hybrid model with indexing allows you to get better forecasting, as a rule, if $\tau < 3$. Data for the experiment were collected by the hardware and software complex for environmental monitoring, which is included in the developed infocommunication system for monitoring the state of the environment for the city of Kiev (Ukraine). On the basis of the developed models, time series of levels of different pollutants were estimated (retrospective data volume up to 1000 points). The general model sets included Holt, Holt-Winters, and medium-wise models with different periods. As a measure of proximity, an Euclidean distance is chosen for the indexation procedure. For example, to predict the level of carbon dioxide pollution in the central part of Kyiv, it was discovered, that the minimum relative error for $\tau \leq 2$ among all the models that function in the system, corresponds to a combined hybrid model with indexing of time series and makes up approx. 1.2%. For $\tau \geq 3$ the minimum is the error of a selective indexing model, and for example for $\tau = 3$ is approx. 2.4%.

Infocommunication system for monitoring the state of the environment allows to keep under observation the weather conditions and microclimate on particular site (in the city, but also in the garden, green house, farm, field) all the year round in real-time mode from any point of the globe. All this, surely help us to monitor the state of the environment, to predict the level of carbon dioxide pollution, etc. The results of the Infocommunication system for monitoring the state of the environment (Inspector Meteo) are shown in Fig. 1.

The system based on the Universal Programmable Logic Controller (see Fig. 2) was used to obtain air pollution and forecasting reports.

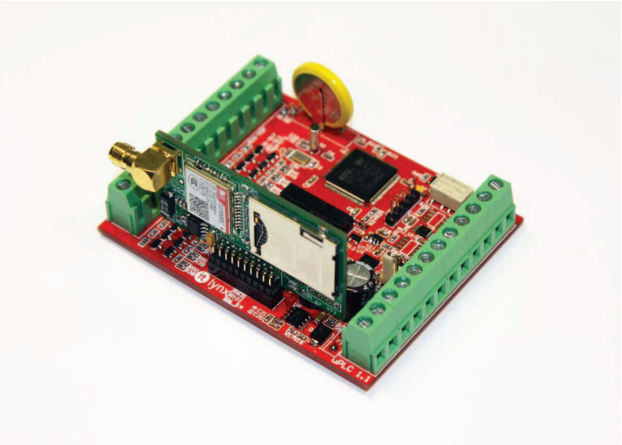

Figure 2. Universal Programmable Logic Controller

Programmable logic controller Universal Programmable Logic Controller modular architecture is designed to solve the problems of telematics, control and management:

- Management of electronic and electro-mechanical equipment;
- Automation and control of technological processes;
- Collection, processing, storage and data transmission.

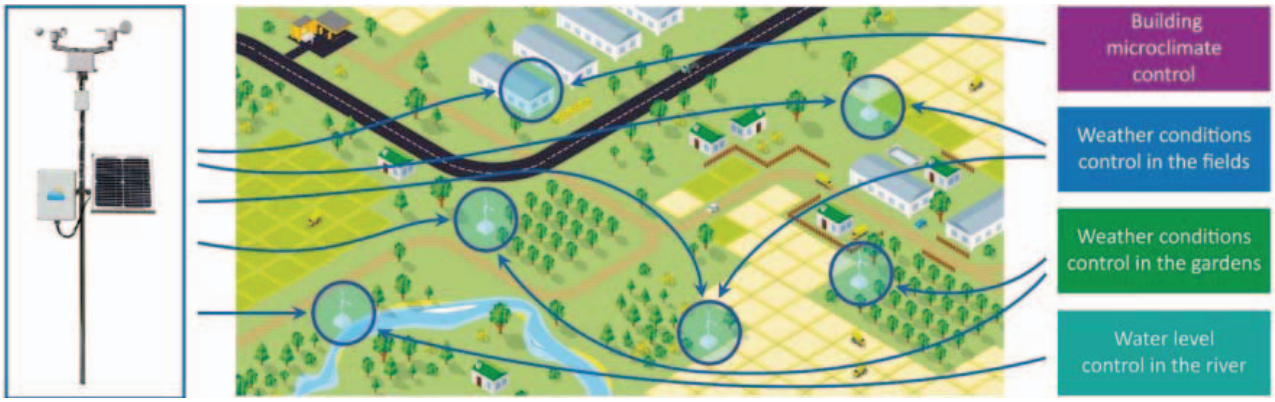

Figure 3. The scope of activities of the Inspector Meteo

The main feature of the controller is the ability to work in different modes: autonomous and with the high level software integration to configure, administrate and manage the controller. Configuration for the controller was developed using integrated Lua programming language interpreter.

In carrying out experiments, we used a GSM module for transmitting measurements results to a web server. Data was compressed before sending. All data processing and pollution forecasting were carried out on the server. The UPLC modular structure allows you to use more energy efficient and cheap solutions based on the IEEE 802.15.4 standard using the Zigbee protocol [17, 18].

Infocommunication system for monitoring the state of the environment (Inspector Meteo), in addition to forecasting environmental pollution, can also perform a number of tasks, such as (see. Fig. 3).

IV. CONCLUSIONS

The combined models of selective and hybrid types with indexation are developed. They can be used to predict the level of air pollution as components of infocommunication systems for monitoring of the environment. These models allow achieving a higher precision of short-term and medium-term forecasting compared to models included in the base set of combined models.

Developed models of prediction of time series can be used not only for the task of forecasting the air pollution level. These models can also be effectively used for the task of forecasting financial, technological and other types of indicators that are represented by time series.

REFERENCES

[1] M. Gardner, and S. R. Dorling, "Neural Network Modelling and Prediction of Hourly NOx and NO2 Concentrations in Urban Air in London," Atmospheric Environment, vol. 5, no. 33, pp. 709-719, 1999.

[2] M. Kolehmainen, H. Martikainen, and J. Ruuskanen, "Neural Networks and Periodic Components Used in Air Quality Forecasting," Atmospheric Environment, vol. 5, no. 35, pp. 815-825, 2001.

[3] Chi-Man Vong, Ip Weng-Fai, Wong Pak-Kin, and Yang Jing-Yi, "Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines," Journal of Control Science and Engineering 2012, pp. 1-11, 2012.

[4] L. C. Ochando, C. I. F. Julian, F. C. Ochando, and C. Ferri, "Airvlc: An application for real-time forecasting urban air pollution," Proceedings of the 2-nd International Workshop on Mining Urban Data, Lille, France, pp. 72-79, 2015.

[5] K. B. Shaban, A. Kadri, and E. Rezk, "Urban Air Pollution Monitoring System With Forecasting Models," IEEE Sensors Journal, vol. 16, no. 8, pp. 2598-2606, 2016.

[6] M. A. Elangasinghe, N. Singhal, K. N. Dirks, and J. A. Salmond, "Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis," Atmospheric Pollution Research, vol. 5, no. 4, pp. 696-708, 2014.

[7] D. Q. Goldin, and P. C. Kanellakis, "On Similarity Queries for Time-Series Data: Constraint Specification and Implementation," 1995 1-st International Conference on the Principles and Practice of Constraint Programming: Cassis, France, pp. 137-153, 1995.

[8] F. Fernández-Rodríguez, S. Sosvilla-Rivero, and J. Andrada-Félix, "Nearest-neighbour predictions in foreign exchange markets," Fundacion de Estudios de Economia Aplicada, no. 5, pp. 1-36, 2002.

[9] M.S. Perlin, "Nearest neighbor method," Revista Eletrônica de Administração, vol. 13, no. 2, pp. 1-15, 2007.

[10] S. Singh, "Pattern modeling in time-series forecasting," Cybernetics and Systems. An International Journal, vol. 31, no. 1, pp. 49-65, 2000.

[11] A. Kuchansky, and A. Biloshchytskyi, "Selective pattern matching method for time-series forecasting," Eastern-European Journal of Enterprise Technologies, vol. 6, no. 4 (78), pp. 13-18, 2015. doi: 10.15587/1729-4061.2015.54812/

[12] A. Berzlev, "A method of increments sings forecasting of time series," Eastern-European Journal of Enterprise Technologies, vol. 2, no. 4 (62), pp. 8-11, 2013. (In Ukrainian)

[13] A. Biloshchytskyi, A. Kuchansky, Yu. Andrashko, S. Biloshchytska, A. Dubnytska, and V. Vatskel, "The Method of the Scientific Directions Potential Forecasting in Infocommunication Systems of an Assessment of the Research Activity Result," 2017 IEEE 4-th International Scientific-Practical Conference "Problems of Infocommunications Science and Technology": Kharkiv, Ukraine, pp. 69-72, 2017.

[14] O. Mulesa, and F. Geche, "Designing fuzzy expert methods of numeric evaluation of an object for the problems of forecasting," Eastern-European Journal of Enterprise Technologies, vol. 3, no. 4 (81), pp. 37-43, 2016.

[15] O. Mulesa, F. Geche, A. Batyuk, and V. Buchok, "Development of combined information technology for time series prediction," Advances in intelligent systems and computing, vol. 689, pp. 361-373, 2018.

[16] A. Biloshchytskyi, A. Kuchansky, Yu. Andrashko, S. Biloshchytska, Ye. Shabala, and O. Myronov, "Development of adaptive combined models for predicting time series based on similarity identification," Eastern-European Journal of Enterprise Technologies, vol. 1, no. 4 (91), pp. 32-42, 2018.

[17] Jong-Won Kwon, Yong-Man Park, Sang-Jun Koo, and Hiesik Kim, "Design of air pollution monitoring wystem using ZigBee networks for ubiquitous-city," 2007 International Conference on Convergence Information Technology, pp. 1024-1031, 2007.

[18] Vasim K. Ustad, A. S. Mali, and Suhas S. Kibile, "Zigbee based wireless air pollution monitoring system using low cost and energy efficient sensors," International Journal of Engineering Trends and Technology, vol. 10, no. 4, pp. 456-460, 2014.