# Effective and Efficient Method for Web   System Performance Forecasting

## Shraddha B.Rahane[1]

[1]*Department of CSE, Pune University, India*

***Abstract:*** *Recent development of Distributed computer systems (DCSs) in networked industrial and manufacturing applications on the World Wide Web (WWW) platform, including service-oriented architecture and Web of Things QoS-aware systems, it has become important to predict the Web performance prediction in time and in space .In this paper, we present Web performance prediction in time by making a forecast of a Web resource downloading using the Turning Bands (TB) geostatistical simulation method. Real-life data for the research were obtained from our own website named "Distributed forecasting system". We generate log file form our website and performing monitoring of a group of Web clients from connected LAN. For better web forecasting we used spatio temporal prediction method with time utility for downloading particular file from website with locations. Also perform comparison study in between spatio temporal method & TB methods. The results show good quality of Web performance prediction using the TB method.*

***Keyword****: Distributed Computer System ,Turning Band,, Qos-Aware Systems, Geostatistics, Spatio-Temporal Prediction.*

## I.    Introduction

New visions of Internet-based distributed computer systems (DCSs) include, among others, engineering industrial ecosystems , networked control systems (NCSs), and service-oriented architectures describe general architecture,   Distributed computer system developed in the web have to deal with communication issues between distributed computers.  Communication issues depend on the specific distributed automation application. The  most important things for such systems are Quality of service (QoS), and security performance reliability. The Web-based Distributed Computer systems are built from DCS nodes which include either Web clients or Web servers (Web services) or both. In client-to-server communication, it is a common situation when a client has to choose a server in case of data or service replication on many servers.  The development of industrial informatics along the five years since the 1st INDIN Conference was marked by most tumultuous transformations in the information and communication technologies (ICT) domain, which are radically and rapidly changing our world. In distributed manufacturing systems, especially in a manufacturing grid (MGrid) system, there are primarily two kinds of manufacturing tasks (or resource service requests): single resource service request task (SRSRTask), which can be completed by invoking only one resource service, and multi-resource service request task (MRSRTask), which is completed by invoking several resource services in a certain sequence. Networked control systems (NCSs) are digital control systems in which the functionality of the sensor, control, and actuator reside in physically different computer nodes communicating over a network. However, random delays and data loss of the communication network can endanger the stability of an NCS. There is increasing evidence, in literature, that the estimation of the electric power quality requires the simultaneous measurement of several quantities and indices in all lines connected to the same point of common coupling.

The increase in the performance that the measuring systems based on digital signal processing techniques has undergone during recent years and the capability of the digital systems of interconnecting and exchanging data are making these systems more and more appealing and cost-effective for power quality applications. Moreover, the availability of a world-wide, low-cost, and public-domain interconnection system, the Internet, is pushing the evolution of the remote measurement systems, where the measurement results. provided by in-field measurement systems are collected and stored by a central unit, toward the distributed measurement systems, where different systems, located in different places, share the same data in order to perform a measurement.  It is known that the major drawback of these systems is the lack of synchronization of the shared data, due to the variable and unpredictable throughput of the net, which may affect the uncertainty of the result of the measurement in a quite significant way.Internet-based distributed computer systems (DCSs) include, among others, engineering industrial ecosystems [1], networked control systems (NCSs) [2], and service-oriented architectures [3], [4]. The Web ecosystem will be a desirable medium for networked platforms for smart things (e.g., sensor networks and embedded devices), as well as for industrial and manufacturing systems. because of supporting the best-effort traffic rather than real time traffic, DCSs developed in the Internet have to deal with challenging communication issues, because of these systems' networking requirements

following the real-time mode of operation. The most important requirements concern performance, reliability, quality of service (QoS), and security [6]. Communication issues depend on the specific distributed automation application, Internet and Web technologies, and character of the network (i.e., private or public, factory field bus, local or wide-area, or wired or wireless).The Web-based DCSs are built from DCS nodes which  include either Web clients or Web servers (Web services) or both. In client-to-server communication, it is a common situation when a client has to choose a server in case of data or service replication on many servers. To make this operation predictable, Web-based DCSs need a spatio-temporal Web performance forecasting, where the temporal prediction addresses the forecasting of Web performance, ordered by one or more dimensions of time (e.g., the minute-by-minute, hour-by-hour, and day-by-day), whereas *spatial* prediction concerns forecasting of the performance behaviour across nodes of DCS (more specifically, between Web clients and Web servers).The aim of this paper is to present a robust spatio-temporal prediction method and algorithm that can provide an efficient forecasting of a Web client-perceived performance on the World Wide Web. This may provide efficient QoS for individual nodes of Web-based DCS and enable to improve operation of the whole system. The predicted performance characteristics can be used in selection of the best performance Web server and best in space and in time. Here, we propose to make Web performance prediction with the use of the Turning Bands (TB) geostatistical method [7].The first is the introduction of a new spatio-temporal methodological approach to the performance prediction of Internet based DCSs, established on the theory and application of geostatistics.

The second is a Web performance prediction algorithm based on the widely proven TB simulation method, which gives efficient and accurate forecasting, as well as reliable results. The third contribution is that our analysis uses real-life data collections gathered for various clients (agents) monitoring many Web servers localized in different Internet geographic locations. At present, to the best of the authors' knowledge, the approach presented in this paper is unique, and there is no similar problem statement in the literature (excluding the authors' research) with which to compare. In particular, any known work deals with the spatio-temporal prediction problem in relation to the Web client-perceived performance. We present the comparison of our TB-based Web performance prediction method with other spatio-temporal prediction approaches, which, like the TB method.

## II.    Literature Survey
In this section we are presenting the different methods those are presented to solve inaccurate performance prediction methods.

- H. Wackernagel[7]*,* presents another geostatistical method is sequential Gaussian simulation (SGS), with the main difference between it and the TB method being the possibility of utilizing sequential neighborhoods to perform a prognosis. These two methods have been used by us because they use an acceptable amount of casts. Geostatistical methods are developing significantly in traditional sciences for geostatistics like climate studies, geology, ecology, or agriculture.
- Kaminska-Chuchmala and A. Wilczynski describes system[22,24] where very promising results were obtained in studying load forecasting in power distribution grids, where generally forecast in transmission networks is a complex task.
- M.Mirza, J. Sommers, P. Barford, and X. Zhu[12,14], explains performance prediction is one of the key.issues of the Internet and the Web. Many works considered the Internet performance through the transmission delay (RTT), data throughput (TCP throughput). Web performance can be evaluated by the Web page view response time and Web resource download time (or HTTP throughput). Both short term and long-term performance forecasts are considered and aimed at either arbitrary or specific file sizes. Short-term forecasting needs instantaneous measuring of network performance that may cause too much additional traffic, therefore lower sampling rates are proposed but for long-term forecasting.
- L. Borzemski and A. Kaminska-Chuchmala present The work in [19] presents the TB method applied to the dataset containing measurements made by the Wroclaw agent that monitored. Web servers dispersed across the world; this paper deals with the application of the TB method for the dataset collected by the agent monitoring European servers from Gliwice.
- Kaminska-Chuchmala and A. Wilczynski persent the work in [12] employs the TB method in the analysis of a Wroclaw agent monitoring European servers.
- Cagdas Hakan Aladag1, Erol Egrioglu the work in [2] uses the Sequential Gaussian Simulation method deployed for Gliwice agent and European servers introduces a new proposal of the spatio-temporal Web performance forecasting.

The survey of forecasting Current forecasting technologies refer to quantitative and qualitative methods. Among the quantitative methods, time-series forecasting methods are used to analyze time-dependent series data and predict the future values, brand choice models are used to calculate the probability of choice to predict choice behavior, and Bayesian models are used to infer conditional probability. Qualitative forecasting technology can be described by environment scanning, scenarios, and Delphi. A review of recent advances in technological forecasting can be found in Martino [19]. The forecasting technologies in the presented architecture focus on quantitative methods.

A survey of time-series data mining was presented by Keogh & Kasetty [10]. Although some accomplishment have been achieved by the effort, in general difficulties still remain because uncertain demand variations do not follow the same pattern all the time due to consumer behavior, new technologies or products, and other environmental factors. Exploring the composite effects of these factors that influence customer demand may capture the future trend and help reduce the errors that might exist in classical time-series forecasting models.

## III.    Related  Work

The TB forecasting method has been originally employed in spatial applied statistics (geostatistics), which is a branch of statistics domain based on spatial data. Its methods can be used for prediction of spatially distributed quantities in diverse disciplines. As a result, one could obtain spatial-temporal database and raster map, where the analyses of variability for whole space not only for given points could be performed. One of the most popular geostatistical methods is the Turning Bands method, which is used, for example, in oceanography and electrical power engineering. Another geostatistical method is sequential Gaussian simulation (SGS) with the main difference between it and the TB method being the possibility of utilizing sequential neighborhoods to perform a prognosis. These two methods have been used by us because they use an acceptable amount of casts. Geostatistical methods are developing significantly in traditional sciences for geostatistics like climate studies, geology, ecology, or agriculture in addition to these, where very promising results were obtained in studying load forecasting in power distribution grids, where generally forecast in transmission networks is a complex task .Performance prediction is one of the key issues of the Internet and the Web. Many works considered the Internet performance through the transmission delay (RTT), data throughput (TCP throughput). Web performance can be evaluated by the Web page view response time and Web resource download time (or HTTP throughput). Both short-term and long-term performance forecasts are considered and aimed at either arbitrary or specific sizes. Short-term forecasting needs instantaneous measuring of network performance that may cause too much additional traffic, therefore lower sampling rates are proposed but for long-term forecasting.

Prediction can be formula-based or history-based Formula-based methods use a mathematical formula ex-pressing particular performance measure as a function of essential independent variables that characterize a studied phenomenon. In history -based performance prediction, the time series of observations obtained through repeated measurements over time are analyzed, and this is the approach used in this paper. Two basic prediction approaches are considered, namely clas-sification and regression. Classification predicts a class from the set of de fined classes whereas the regression predict a numeric value of particular performance measure. Both approaches are in use Machine learning techniques and data mining methods and tools show a good empirical performance for TCP throughput prediction in many applica-tion domains, for example in adaptive applications, such as real-time multimedia streaming applications, which are increasingly using TCP as their transport protocol.

## IV.  PROPOSED APPROACH FRAMEWORK AND DESIGN
### 4.1 Problem Definition
Improving and forecasting the distributed web-system performance using efficient Turning band method.

we have studied the many methods those are presented for forecasting of web system performance by considering the different real time applications.

In [1] we have studied the TB geostastical performance prediction method for distributed computer systems and web systems. Following flowchart with reference this is presented in figure 1.
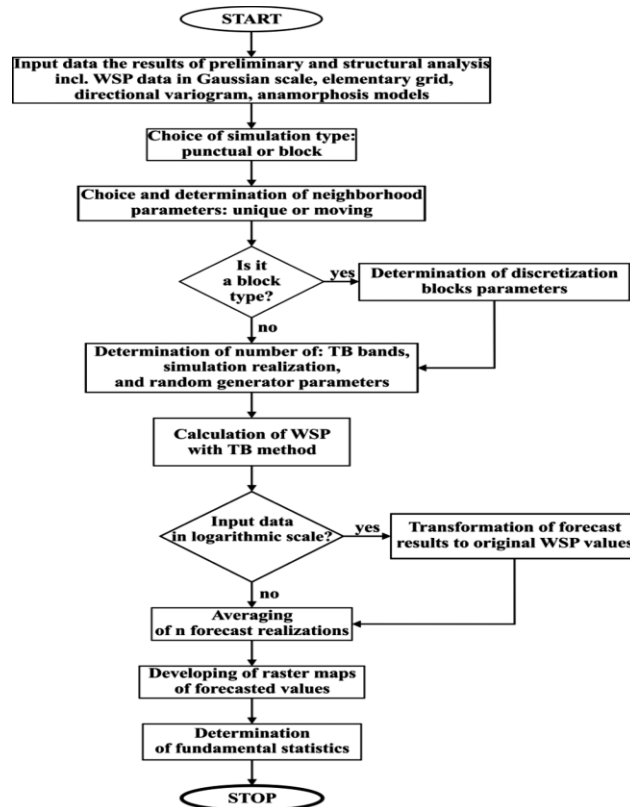
**Figure 1:  TB based flowchart for forecasting performance**

## 4.2 Aims and Objectives

In this project we have main aim is to present the extended method for Distributed Web Systems Performance Forecasting Using Turning Bands Method with improved reliability and performance:

- To present literature review different methods Distributed Web Systems Performance Forecasting Using Turning Bands Method.
- To present the present new framework and methods.
- To present the practical simulation of proposed algorithms and evaluate its performances.
- To present the comparative analysis of existing and proposed algorithms in order to claim the efficiency.

## 4.3 Proposed System Architecture and Algorithms

In this we are presenting the new architecture which is based on two concepts first is spatio temporal prediction (SPT) which depicted in [1] and another one is TB which is depicted in [2]. Based on these terminologies below is proposed architecture for predicting the performance in figure 2.

As shown in figure 2, we first collect log data, then apply preprocessing steps like data cleaning, in cleaning process we remove records where status code is 404 or URL contains images like jpg, png etc. Then apply spatio temporal prediction & TB algorithm, finally we will perform analysis with comparison study. The main idea of the TB method is to perform simulations along several lines, using a unidimensional covariance function that corresponds to the given three-dimensional (3-D) one: X, Y directions—geographical coordinates and Z direction—time axis. The TB algorithm is calculated in cubic time complexity. At each point of the 3-D field, a weighted sum of the corresponding values of the line processes is assigned. In conclusion, the TB method is a multidimensional random number generator for the simulation of the spatially correlated random fields [13], [14].
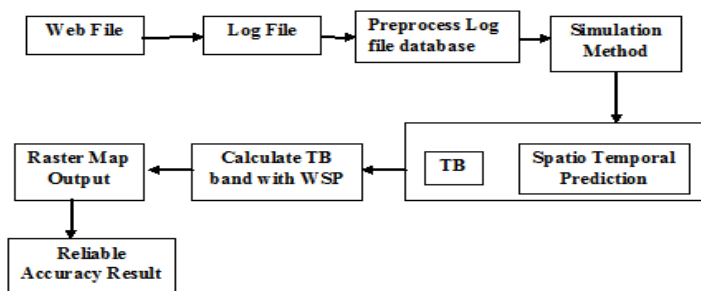
**System Architecture**



**Figure 2: Proposed architecture for predicting the performance**

### 4.4  Turning Bands

The amount of traffic generated on the Internet continuous to grow. In the nineties, only a few households were connected to the network. This situation, however, has changed since the Internet has become ubiquitous. Not only have modern households more than one network-connected device, but companies also commenced to support BYOD model (bring your own device). This brings about considerable number ofdevices generating enormous network traffic (especially in the Web) and makes IT administrators snowed under with their job. To deliver quality-based services, administrators not only need to constantly monitor their resources, but also try to predict possible situations. This paper presents geostatistical approaches, namely Turning Bands and Sequential Gaussian Simulation as a solution to that problem, because these methods have already proven themselves in computer science. These approaches allow to make a spatio-temporal forecasts using only historical data, gathered during daily servers monitoring.To put in a nutshell, we have collected the data concerning performance of fifteen evaluated European web servers, between 06 and 28 February, 2009. A server's performance in this case corresponds to the time required to download the resource from the server. At the outset, these data were subjected to thorough analyses; subsequently, they were used to forecast evaluated web servers' performance from 1 to 4 March, 2009.

Here, we present the methodology of the proposed approach and the algorithm of the TB method, which will  be used for spatio-temporal forecasting of Web system performance (WSP).The basic assumption of the TB method is as follows: the field to be simulated is second-order stationary and isotropic; at each point, the values of the field are normally distributed and have zero mean. In other cases, the transformation to Gaussian with subsequent subtraction of the mean could be applied. The next assumption is the knowledge of the covariance C(r) of the field which is to be simulated. The main idea of the TB method is to perform simulations along several lines, using a uni dimensional covariance function that corresponds to the given three-dimensional (3-D) one: X, Y directions—geographical coordinates; and Z direction—time axis. The TB algorithm is calculated in cubic time complexity. At each point of the 3-D field, a weighted sum of the corre-sponding values of the line processes is assigned. In conclusion, the TB method is a multidimensional random number generator for the simulation of the spatially correlated random fields.

Let ($\theta$n, n £ N) be a sequence of directions $S + d$   and let $(Xn, n \ £ \ N)$ be a sequence of independenta stochastic process of covariance. The  algorithm of TB method takes the following from C$\theta$n.

TB Algorithm:

1. Directions Selection: $\theta_1 \ldots \theta_n$ so that $1/n \sum_{k=1}^{n} \delta \theta_k$ is weakly convergent to $\varpi$

$$\theta_n = (\cos(2\Pi un) \sqrt{1 - v^{\wedge}2n}, \sin(2\Pi un) \sqrt{1 - v^{\wedge}2n} , vn).$$

Un= binary expansion of each integer n=1,2,……………;
Vn= ternary expansion of each integer n=1,2……………..;

2. Convariance function

$$C (n)(h) = 1/n + \sum_{k=1}^{n} (C\theta k \ (< h, \theta k >)) \ \sim \varpi$$

$$\theta_{n=} \theta_k$$

3. Calculate x Ɛ D

$$1/sqrt(n) + \sum_{k=1}^{n} (C\theta k \ (< h, \theta k >))$$

X= forecasted download time;

## 4.5  Spatial  Prediction

The Formula-based methods use a mathematical formula expressing particular performance measure as a function of essential independent variables that characterize a studied phenomenon. In history-based performance prediction, the time series of observations obtained through repeated measurements over time are analyzed, and this is the approach used in this research.

**Spatial Prediction Algorithm:**

For each user we keep track of all previous time taken to download file from particular ip address. The algorithm predicts the next time taken to download file to a given location by means of the previous history of time

$$((t1, d1), (t2, d2),\ldots\ldots\ldots\ldots (tn, dn));$$

1. Two time series are created from the sequence of previous time series: the time series of the visit daily start times C and the time series of the visit durations D defined as follows:

$$C = (c1, c2,\ldots\ldots\ldots, cn)$$
$$D = (d1, d2,\ldots\ldots\ldots.., dn)$$

Where $c_i$ is the time of the day in seconds corresponding to the time instant $t_i$ (i.e., $c_i$ is in the range [0; 86400]);

2. we search in the time series C sequences of m consecutive values (ci-m+1; : : : ; ci) that are closely similar to the last m values (cn-m+1; : : : : ; cn);

3. The next value of time series C is estimated by averaging all the values ci+1 that follow each found sequence;

4. At the same time, in time series D the corresponding sequences (di-m+1; : : : ; di) are selected; the sequences have to be located exactly at the same indexes as those in C;

5. The next value of time series D is then estimated by averaging the entire values di+1 that follow these sequences.

## V.    Experimental Setup

The Web performance measurement database was created during active monitoring experiments conducted by the multi-agent system MWING; its agents were installed on local hosts in academic campuses in four geographical locations that collect information on various Internet and WWW characteristics. MWING uses local as well as a centralized database, therefore locally gathered data can be uploaded in one place allowing further analysis of data, e.g., with the use of a data mining engine. MWING supports various types of agents implemented in different programming languages, so it can be run in both Linux and Windows operating environments. Agents perform measurements and monitoring by means of common system functionalities as well as on open developments aiming to match specific aims of measurements. Common functionalities include agent management, measurement scheduling, heartbeat (status and conditions of an agent), data model, synchronization, local databases, and central database support. MWING collects all measurements in a central relational database allowing to analyze the data in the form of time series, as each measurement is time stamped and identified by the monitoring and monitored locations. Because of its critical role for the web users, the Web performance should be considered from the perspective of a web client, so it is perceived as the response times on the WWW. From a client's perspective, the response time is the elapsed time from the initiation of a web page download request at a client side to the time that the response is fully loaded by that client. MWING collects all data about time intervals which characterize so-called web transaction (i.e., client-to-server interactions), starting from the Domain Name Service (DNS) interaction and TCP connection establishment, whole web page download including embedded objects, up to TCP connection termination. For each web transaction, 20 attributes were measured. In fact, three main factors contribute to the response time perceived by a web client: the network delay, the web server latency, and the delay caused by the special web infrastructure, built on the client-to-server communication path to reduce the response time, if only exists.

Finally, a web client always perceives the grand total delay resulted from all activities. To monitor client-perceived Web performance, we use the Web browsing method which in some sense is similar to the ping utility. First, an object (file) which is likely to be available on many web servers all over the world. Next, in a preliminary analysis, the web server reachability was checked, and geographic and organizational information about each server was collected. In the experiment performed within the MWING's infrastructure and give accuracy result.

## VI.    Results Of Practical Work

As dataset we use our own web applications log file generated by IIS server, for that we design one web application named as ''distributed web forecasting.

```
<?xml version="1.0" encoding="us-ascii" standalone="yes" ?>
 - <Web-log>
  - <record>
      <Date>2013-11-30</Date>
      <Time>07:04:12</Time>
      <Server_IP>::1</Server_IP>
      <Method>GET</Method>
      <URL>/Web_Performance_Prediction/Home.aspx</URL>
      <URI_Query>-</URI_Query>
      <Port>80</Port>
      <User_Name>-</User_Name>
      <Client_IP>::1</Client_IP>
      <Browser_Used>Mozilla/5.0+(Windows+NT+6.1;+rv:26.0)+Gecko/20100101+Firefox/26.0</Browser_Used>
      <Status>200</Status>
      <Sub_Status>0</Sub_Status>
      <Win_Status>0</Win_Status>
      <Time_Taken>1225</Time_Taken>
    </record>
  </record>
```

**Figure 3: Content of LOG file generated by IIS**

Figure 3 shows some content of Log file which we used for forecasting a download time. It contains Date-Time, server IP address, http method, URL, port number, client IP address, browser used, status code, time taken to load page.As we required only few data, so only extract needed data like client IP, URL, status code, time taken.
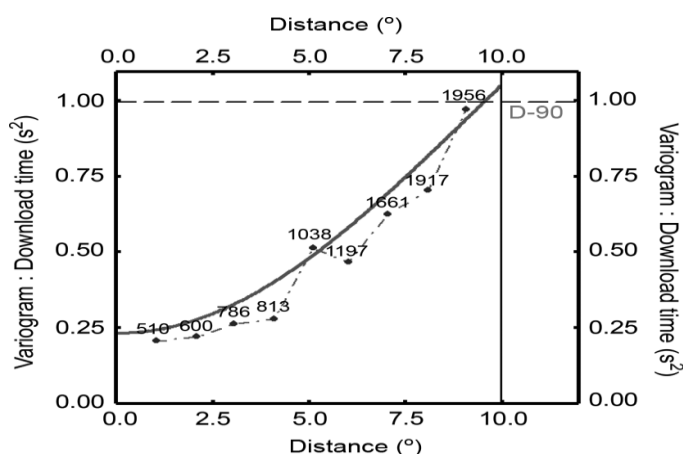


**Figure 4:  Directional variogram along the time axis for Web Performance Prediction Result.**

Figure 4 shows Web performance prediction result using Turning bands algorithm considering the time series and the  downloading time taken this parameter used and shows the variogram result.
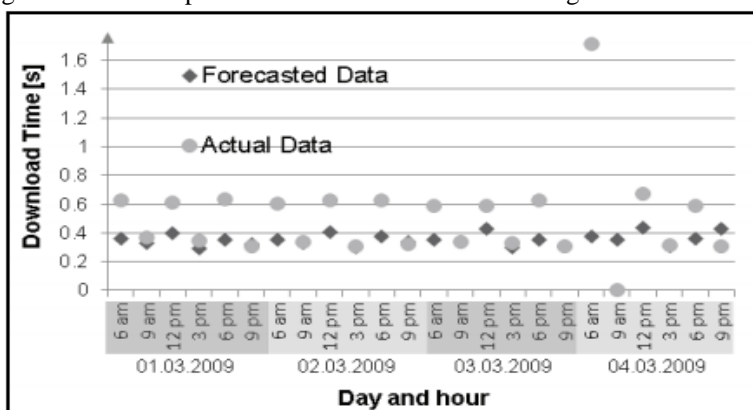


**Figure 5: Actual and forecasted performance using TB**

Figure 5 show the accuracy of the result and calculated the actual forecast result using Spatial prediction method and turning bands algorithms. this result considering the parameter time taken downloading and predictions time to downloading and shows the forecasting result.

## VI.   Conclusion And Future Work

In this research, we implement a TB method to improve the efficiency making daily analyses and spatio-temporal forecasts of web servers' performance. A large-scale measurement experiment performed in the real-life Internet to gather the data characterizing performance of over LAN and perceived from different clients installed in different Internet locations. Such analyses and forecasts may be very helpful for IT administrators, especially in analyzing both network traffic and web servers performance. Future work, there is still need to improve the accuracy forecast this hybrid method and present approach which can handle the different kinds of measurements and data. As well improve the proposed results and security.

## Acknowledgements

## References

[1]     A. Kaminska-Chuchmala and A. Wilczynski, "Application of simulation methods to spatial electric load forecasting," *Rynek Energii*, vol. 80, no. 1, pp. 2–9, 2009.

[2]     A. Kaminska-Chuchmala and A. Wilczynski, "Spatial electric load forecasting in transmission networks with sequential Gaussian simulation method," *Rynek Energii*, vol. 92, no. 1, pp. 35–40, 2011.

[3]     A. Kaminska-Chuchmala and A. Wilczynski, "Analysis of different methodological factors on accuracy of spatial electric load forecast performed with turning bands method," *Rynek Energii*, vol. 87, no. 2, pp. 54–59, 2010

[4]     M. Ulieru and S. Grobbelaar, "Engineering industrial ecosystems in a networked world," in *Proc. 5th Int. IEEE Conf. Ind. Informat.*, Vienna, Austria, Jul. 23–27, 2007, keynote address.

[5]     *Internet-based Control Systems: Design and Applications, Advances in Industrial Control*, S-H. Yang, Ed. London, U.K.: Springer-Verlag, 2011.

[6]     F. Tao, D. Zhao, Y. Hu, and Z. Zhou, "Resource service composition and its optimal-selection based on particle swarm optimization in manufacturing grid system," *IEEE Trans. Ind. Inform.*, vol. 4, no. 4, pp. 315–327, Nov. 2008.

[7]     T. Cucinotta, A. Mancina, G. F. Anastasi, G. Lipari, L. Mangeruca, R. Checcozzo, and F. Rusina, "A real-time service-oriented architecture for industrial automation," *IEEE Trans. Ind. Inform.*, vol. 5, no. 3, pp.267–277, Aug. 2009.

[8]     D. Guinard, V. Trifa, F. Mattern, and E. Wilde, "From the Internet of things to the web of things: Resource oriented architecture and best practices," in *Architecting the Internet of Things*, D. Uckelmann, M. Harrison, and F. Michahelles, Eds. Berlin, Germany: Springer, 2011, pp. 97–129.

[9]     Borzemski l., cichocki l., kliber m., fras m., nowak z., mwing: a multiagent  system for web site measurements, in: lecture notes in computer science, 4496, 2002, 278–287.

[10]    Borzemski l., cichocki l., kliber m., a distributed system to measure the internet based  on agent architecture, in: information systems architecture and technology, web-age information systems, eds leszek borzemski [et al.], oficyna wydawnicza politechniki wrocławskiej, 2009.

[11]    Borzemski l., nowak z., empirical web performance evaluation with using a mwing system, in: information systems architecture and technology: advances in web-age information systems, eds: leszek borzemski [et al.], oficyna wydawnicza politechniki wroclawskiej, 2009, 25–34.

[12]    Lee Y.S., Tong L.I., "Forecasting time series using a me-thodology based on autoregressive integrated moving average and genetic programming," Knowledge-Based Systems 24(1) (2011) pp:66-72.

[13]    L. Borzemski, L. Cichocki, andM. Kliber, "Architecture of multiagent Internet measurement system MWING release 2," *LNCS*, vol. 5559, pp. 410–419, 2009.

[14]    L. Borzemski, "The experimental design for data mining to discover web performance issues in a wide area network," *Cyber. Syst.*, vol. 41, no. 1, pp. 31–45, 2010.

[15]    Y.-M. Lee and J. H. Ellis, "Turning bands method and its application to geostatistical analysis of groundwater contaminant concentration fields," *J. Chin. Inst. Environ. Eng.*, vol. 10, no. 1, pp. 1–12, 2000.