# Risk Prediction of the SCADA Communication Network Based on Entropy-Gray Model

Meng Li, Wenjing Li, Peng Yu and Fanqin Zhou
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, China
Beijing, 100876, P. R. China
Email: lidameng@bupt.edu.cn

***Abstract*—The power SCADA system is designed to ensure the safe operation of the power system. The SCADA communication network as an information exchange carrier between remote terminal units and master stations, is the key part of the SCADA system, and it has a high requirement for security. However, due to the wide distribution of the network and the interconnected network structure, it is susceptible to risks. So there is an urgent need for accurate and real-time risk prediction. In this paper, we propose a risk prediction model based on entropy-gray model, where the gray model is used to predict the values of the network risk indexes, and the entropy method is to determine the weight of those risk indexes. Finally, the overall risk value of the network is decided with analytic hierarchy process. Simulation results show that the proposed entropy-gray method can achieve accurate and timely risk prediction.**

***Keywords*—SCADA communication network, risk prediction, gray model, entropy method**

## I. Introduction

With the rapid innovation of information technology, the power SCADA system gradually turned from closed to open [1]. It mainly consists of three parts, divided into remote terminal units (RTUs), master stations and the remote communication network, which is responsible for monitoring the running real-time data of the remote device and issuing a control command. SCADA system involves the power generation, substation, transmission, and distribution of all aspects of the power grid, providing reliable protection for the safe and continuous work of the power system.

SCADA communication network as a carrier between remote terminal units and master stations for information exchange, is the key to the normal operation of the SCADA system. In China, SCADA communication network is deployed in the power scheduling data network and in the event of a failure, it may lead to paralysis of the entire power system. Therefore, in order to ensure the safety of the increasingly important power system, it is necessary to change the traditional fault generation and passive response mechanism into an active paradigm and conduct risk assessment and prediction research for the power SCADA communication network.

In this paper, a risk prediction mechanism of the power SCADA communication network based on entropy-gray model is proposed. We take advantages of gray prediction theory for the multi-index prediction of the SCADA communication network and the entropy method to determine the weight of the indexes. In summary, the main contributions of this paper include: 1) A multi-index prediction model of the network based on gray model (GM) is proposed to predict the value of each index that affects the network risk. 2) The method of calculating the weight of each risk index based on entropy theory is proposed, in which weight distribution entirely depends on the actual problem domain rather than manual experience, thus eliminating the subjective randomness.

The rest of this paper is organized as follows. Section II reviews the related work. Section III describes the prediction model. Section IV presents the simulation results and the conclusion is presented in Section V.

## II. Related Work

The problem about the risk of the power SCADA system has already been extensively studied. But most existing research is concentrated on the entire SCADA system risk [2]–[4]. One of the excellent methods is presented in [4], where the risk index hierarchy of power SCADA system is constructed and the hybrid algorithm of Gaussian process is optimized by artificial bee colony algorithm to realize the risk assessment of SCADA system.

Considering the importance and vulnerability of the SCADA communication network, it is necessary to conduct research specifically on the network risks, while the existing studies on it only focus on individual risks [5], [6]. Reference [7] achieves the SCADA communication network accurate modeling through comprehensive analysis of communication networks, but it does not achieve effective risk prediction.

Forecasting techniques for individual risks have been widely studied. Reference [8] uses KL divergence to analyze the risk of power failure, but it is too simple that it does not apply to other risks. In [9], [10], the GM prediction is used to predict the power load. It shows superior performance in different weather conditions. With the rise of machine learning, it is possible to study the overall risk of a complex network. In [11], a risk prediction method based on neural network is proposed for the power grid. Yet it requires a lot of historical data and enough training times, and the neural network parameters need to constantly debug.

In order to realize accurate risk prediction of the SCADA communication network, this paper introduces the gray model

978-3-901882-98-2 

(GM) to achieve multi-index prediction and the entropy method to determine the weight of the risk indexes.

## III. Risk Prediction Based on Entropy-Gray Model

To conduct risk research of the SCADA communication network, we first need to clear the specific process of the power scheduling data network risk warning, generally divided into three stages: risk analysis, risk prediction and risk alarm. It can be seen that accurate risk prediction is the key to network risk warning. In this paper, a risk prediction mechanism based on entropy-gray model is proposed for the power SCADA system. The specific process is shown in Fig. 1.

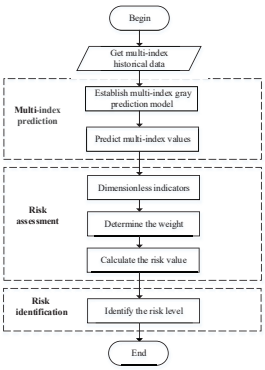


Fig. 1. Process of the risk prediction

### A. Multi-index Prediction Based on Gray Model

The gray prediction model is based on the time series, and the model parameters are updated in real time. By constructing the accumulated sequence of multi-index data, it can enhance the real-time prediction and the anti-jamming of the pathological data. The prediction model is constructed as follows:

*1) Generate an accumulation matrix:* Assuming that there are $m$ indexes affect the risk of the network, each index is sampled $n$ times, $x^{(i)}(j)$ is the index value of the $j$-th sample of the $i$-th index. Let $X^{(i)}(k) = \sum_{j=1}^{k} x^{(i)}(j), k = 1, 2, \cdots, n$, that means $X^{(i)}(k)$ is the cumulative sum of the first $k$ moments of the $i$-th index. Hence the new accumulation matrix is

$$X = \begin{bmatrix} X^{(1)}(1) & X^{(1)}(2) & \cdots & X^{(1)}(n) \\ X^{(2)}(1) & X^{(2)}(2) & \cdots & X^{(2)}(n) \\ \cdots & \cdots & \cdots & \cdots \\ X^{(m)}(1) & X^{(m)}(2) & \cdots & X^{(m)}(n) \end{bmatrix}. \quad (1)$$

*2) Construct the gray differential equation and solve it:* The the following formula is gray differential equation

$$x^{(i)}(k) + aZ^{(i)}(k) = b, \quad (2)$$

here $Z^{(i)}(k) = \frac{1}{2}[X^{(i)}(k) + X^{(i)}(k-1)], k > 1$, and $a, b$ is the undetermined coefficient, known as the development coefficient and the role of gray, respectively. Assuming the matrix $A = \begin{pmatrix} a \\ b \end{pmatrix}$ as the gray parameter, as long as we solve $A$, we can have $X^{(i)}(k)$ and $x^{(i)}(k)$. It can be seen that the gray differential equation is a multivariate regression equation. In order to solve $A$, we can use the least squares principle, as follows

$$\hat{A} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (B^T B)^{-1} B^T Y_n, \quad (3)$$

where the mean generation $B$ and the constant term vector $Y_n$ are denoted as

$$B = \begin{bmatrix} -\frac{1}{2}\left(X^{(i)}(1) + X^{(i)}(2)\right) & 1 \\ -\frac{1}{2}\left(X^{(i)}(2) + X^{(i)}(3)\right) & 1 \\ \cdots & \cdots \\ -\frac{1}{2}\left(X^{(i)}(n-1) + X^{(i)}(n)\right) & 1 \end{bmatrix}$$
$$Y_n = \begin{bmatrix} x^{(i)}(2) \\ x^{(i)}(3) \\ \cdots \\ x^{(i)}(n) \end{bmatrix}. \quad (4)$$

Substitute the calculated results $\hat{a}$ and $\hat{b}$ into (3), we get

$$\hat{X^{(i)}}(k+1) = (X^{(i)}(1) - \frac{\hat{b}}{\hat{a}})e^{-\hat{a}k} + \frac{\hat{b}}{\hat{a}}, k = 1, 2, \ldots n. \quad (5)$$

The actual forecast data can be obtained by the following equation

$$\begin{aligned} \hat{x^{(i)}}(k+1) &= \hat{X^{(i)}}(k+1) - \hat{X^{(i)}}(k) \\ &= (1 - e^{\hat{a}})(x^{(i)}(1) - \frac{\hat{b}}{\hat{a}})e^{-\hat{a}k}, k = 2, 3, \ldots n \end{aligned}. \quad (6)$$

*3) Correct the predicted value:* From (6) we can see that the gray prediction result is an exponential model. Assuming that the original data sequence is an exponential sequence in the form of

$$x^{(i)}(k) = Me^{a(k-1)}, k = 1, 2, \ldots, n. \quad (7)$$

Using the traditional gray prediction model GM (1, 1), the final fitting result is

$$\hat{x^{(i)}}(k) = \frac{-Me^{a}(1 - e^{\hat{a}})}{1 - e^{a}} e^{-\hat{a}(k-1)}, k = 2, 3, \ldots, n. \quad (8)$$

Compare (7) with (8), we can see that the traditional gray prediction model has the deviation.

By observing the above equations, it can be found that we can use $\hat{a}$ and $\hat{b}$ to represent $a$ and $M$ in the original data sequence, the improved GM (1, 1) model parameters are obtained as

$$\tilde{a} = \ln \frac{2 - \hat{a}}{2 + \hat{a}}, \tilde{M} = \frac{2\hat{b}}{2 + \hat{a}}. \quad (9)$$

The final prediction equation is

$$\tilde{x^{(i)}}(k) = \tilde{M}e^{\tilde{a}(k-1)}, k = 2, 3, \ldots n. \quad (10)$$

Thus, the risk index prediction matrix is obtained as

$$\tilde{X} = \left[X^{(1)}(\tilde{k}+1) \quad X^{(2)}(\tilde{k}+1) \cdots X^{(m)}(\tilde{k}+1)\right]^T. \quad (11)$$

The modified predictive value does not have the inherent deviation of the preliminary predictive value and it simplifies the modeling step.

### B. Risk Assessment Based on Entropy Method

*1) The non-dimensional of indexes:* Because of the different dimension of different indexes, it is necessary to dimensionless the risk index matrix to eliminate the incomparability of different dimensions. In this paper, we use the optimal method of extreme value processing [12] and control the quantization results on [0,100].

According to the index attributes, the indexes can be divided into four categories: positive correlation (the greater the value, the higher the risk), negative correlation (the greater the value, the lower the risk), boolean indexes (indexes only have two opposing cases) and fuzzy indexes (indexes can only be described in words, good quality, general, poor, etc.).

On the basis of the classification of indexes, through the following equation for non-dimensional processing:

a) Positive correlation

$$\lambda_k = \frac{x(k) - \min\limits_k\{x(k)\}}{\max\limits_k\{x(k)\} - \min\limits_k\{x(k)\}} \times 100, k = 1, 2, ..., n. \quad (12)$$

b) Negative correlation

$$\lambda_k = \frac{\max\limits_k\{x(k)\} - x(k)}{\max\limits_k\{x(k)\} - \min\limits_k\{x(k)\}} \times 100, k = 1, 2, ..., n. \quad (13)$$

c) Boolean indexes

As some indexes only have two opposing cases, so we set the value of 0 or 100 according to the actual situation.

d) Fuzzy indexes

Remove the fuzzy by expert evaluation, it is converted into a clear value between 0 and 100.

*2) Entropy method to determine the weight:* When the risk assessment is carried out, a particular index of a large difference indicating that it contains more information and it has a stronger ability to distinguish the system, so we should give it a greater weight. In this paper, the entropy method is used to determine the index weight.

The risk index matrix is $x = x^{(i)}(j)_{m\times n}$, of which $x^{(i)}(j)$ is the index value of the $j$-th sample of the $i$-th index, the weight $p_{ij}$ is calculated as follows

$$p_{ij} = \frac{x^{(i)}(j)}{\sum\limits_{j=1}^{n} x^{(i)}(j)}. \quad (14)$$

Calculate the entropy of the $i$-th index

$$e_i = \frac{1}{\ln n} \sum_{j=1}^{n} p_{ij} \ln \frac{1}{p_{ij}}. \quad (15)$$

The difference coefficient of the $i$-th index is

$$g_i = 1 - e_i. \quad (16)$$

Then the $i$-th index weight is

$$\omega_i = \frac{g_{\mathrm{i}}}{\sum\limits_{i=1}^{m} g_i}. \quad (17)$$

*3) Calculate the risk value of the network:* The value is calculated as follows

$$R_N = \sum_{i=1}^{m} (\lambda_i \times \omega_i), \quad (18)$$

where $R_N$ is the network risk value, $\lambda_i$ is the predictive value of the $i$-th risk index and $\omega_i$ is the weight of the $i$-th index.

The risk of the whole network can be measured mainly from two aspects: on the one hand, the network risk value; on the other hand, the network risk discrepancy, which is used to measure the degree of dispersion of the risk value in the network. It is calculated by the following equation

$$V_N = \frac{1}{R_N} \sqrt{\sum_{i=1}^{m} [\omega_i \times (\lambda_i - R_N)^2]}, \quad (19)$$

where the network risk discrepancy is $V_N$, the network risk value is $R_N$, the risk value of the $i$-th index is $\lambda_i$ and the weight of the $i$-th index is $\omega_i$. The greater the risk discrepancy is, indicating that the network operation is more unreasonable.

### C. Identification of Network Risk Level

By combining the network risk value and the degree of the network risk discrepancy, we can get the risk level of the network. The correspondence between the above three is shown in Table 1.

TABLE I
RELATIONSHIP BETWEEN NETWORK RISK LEVEL AND RISK VALUE AND RISK DISCREPANCY

| Risk Level | Description | $R_N$ and $V_N$ |
|---|---|---|
| 5 | Critical | $(\beta_4, 100]$ *or* $(\eta_4, \eta_5]$ |
| 4 | High risk | $(\beta_3, \beta_4]$ *or* $(\eta_3, \eta_4]$ |
| 3 | Medium risk | $(\beta_2, \beta_3]$ *or* $(\eta_2, \eta_3]$ |
| 2 | Low risk | $(\beta_1, \beta_2]$ *or* $(\eta_1, \eta_2]$ |
| 1 | Safe | $(0, \beta_1]$ *or* $(0, \eta_1]$ |

Where the interval $(\beta_i, \beta_{i+1}]$ and $(\eta_i, \eta_{i+1}]$ respectively represent the ranges of $R_N$ and $V_N$, corresponding to the risk level, and the specific values are given according to the actual situation of the network (such as different requirements of safety and real-time).

## IV. Experimental Results and Analysis

### A. Experimental Data

A lot of risk factors can affect the safety of the network in the actual environment. Based on the principle of comprehensiveness and scientific, this paper selects three categories of 20 indexes for simulation experiments.

The experimental data is determined according to the actual operation of the power scheduling data network in 2016. In this paper, neural network method, fixed weight method (that is, only gray multi-index prediction, the weight of each index is fixed) and the method of directly using network risk value to predict (that is, using the assessed network risk value to predict) are chosen to compare with the entropy-gray prediction method proposed in this paper, the experiment is simulated by MATLAB (2016b).

### B. Experimental Results and Analysis

The accuracy of the algorithm is measured by the mean relative error (*MRE*), the root mean square error (*RMSE*) and the mean absolute error (*MAE*).

$$MRE = \frac{1}{S}\sum_{l=1}^{S}\left[\frac{(R_N^l - R_N^{l'})}{R_N^l}\right] \times 100\%, \tag{20}$$

$$RMSE = \sqrt{\frac{\sum_{l=1}^{S}(R_N^l - R_N^{l'})^2}{S}}, \tag{21}$$

$$MAE = \frac{1}{S}\sum_{l=1}^{S}\left|R_N^l - R_N^{l'}\right|, \tag{22}$$

where $S$ is the number of experimental samples, $R_N^l$ is the actual risk value of the network for the $l$-th sample, and $R_N^{l'}$ is the predicted risk value of the network for the $l$-th sample.

Fig. 2 shows the prediction results of the proposed entropy-gray method, the neural network method, the fixed weight method and the direct use of network risk value prediction method.

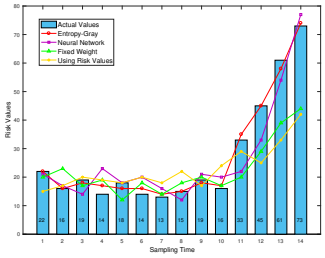


Fig. 2. Results of different prediction methods

Fig. 3 shows the error analysis for each prediction method. The method of using the network risk value to predict shows the worst performance, and sometimes even predicts the opposite trend with the actual network risk, since it does not consider the current network operating state. The method of fixed index weight is less responsive when the network risks change dynamically. This is because it has no consideration that with the dynamic changes in the network, the impact of different risk indexes also changes with dynamic. The prediction accuracy of BP neural network is better, but the trend to follow up is obviously worse than the entropy-gray prediction. This is because in the dynamic network, the available historical data are limited, the neural network cannot reach enough training times, so generalization ability is insufficient. While the entropy-gray model only considers a small amount of data before the forecast time. Not only the accuracy is guaranteed, but the time and space complexity of the algorithm is much lower than that of the neural network.

Above all are only in the comparison of the network risk value, in order to achieve a more accurate risk level, we need to consider the degree of network risk discrepancy, just as

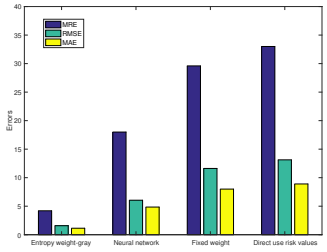


Fig. 3. Comparison of algorithm error results

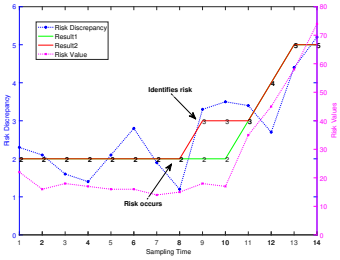


Fig. 4. Comparison of discriminant results

shown in Fig. 4. The two dashed lines are the values of the risk value and the risk discrepancy. The two solid lines represent the discriminant results of the risk level. The green one only considers the risk value. The red result takes the risk value and the risk discrepancy into account, which can give a higher risk alarm earlier. This is because the degree of the network risk discrepancy describes the degree of discrepancy between each risk indexes, which can identify the network risk caused by a certain index. In summary, the simulation results validate that the entropy-gray model has better performance for the risk prediction of the SCADA communication network.

## V. Conclusion

In order to achieve risk prediction of power SCADA communication network, this paper proposes a risk prediction model based on entropy-gray model. By using the gray prediction model based on time series, the prediction accuracy under the dynamic changes of the network is ensured. Moreover, the entropy method is used to characterize the dynamic evolution of those risk indexes, and the weight of each index depends entirely on actual problems rather than manual experience. Finally, we jointly consider the network risk value and the network discrepancy to judge the risk level, and reasonable alarms can be generated accordingly. In the future, we plan to improve the proposed approach by introducing the correlation analysis of risk indexes to achieve a more accurate risk prediction and assessment.

## Acknowledgment

This work was supported by State Grid technical project (52010116000W), National Natural Science Foundation of China (NSFC) (Grant No. 61672108) and Tibet Natural Science Foundation (Grant No. 2016ZR-15-63).

## REFERENCES


[1] Y. Zhang and J. L. Chen, "Wide-area scada system with distributed security framework," *Journal of Communications and Networks*, vol. 14, no. 6, pp. 597–605, 2012.

[2] M. B. S. Kriaa and Y. Laarouchi, "A model based approach for scada safety and security joint modelling: S-cube," in *The IET System Safety and Cyber Security*, 2015, pp. 1–6.

[3] P. Z. S. Katam and F. Gichohi, "Applicability of domain based security risk modeling to scada systems," in *World Congress on Industrial Control Systems Security*, 2015, pp. 66–69.

[4] C. Shengnan, "Power scada system vulnerability analysis and network security risk assessment research," Ph.D. dissertation, North China Electric Power University, 2015.

[5] P. M. G. K. Chalamasetty and T.-L. Tseng, "Secure scada communication network for detecting and preventing cyber-attacks on power systems," in *Power Systems Conference*, 2016, pp. 1–7.

[6] Y. X. Y. Zhang, L. Wang and C. W. Ten, "Inclusion of scada cyber vulnerability in power system reliability assessment considering optimal resources allocation," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4379–4394, 2016.

[7] K. M. Z. C. C. H. A. Hou, C. Hu and T. Pan, "Research on modeling and simulation of communication in power scada system," in *International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*, 2016, pp. 226–230.

[8] S. Gupta, S. Waghmare, F. Kazi, S. Wagh, and N. Singh, "Blackout risk analysis in smart grid wampac system using kl divergence approach," in *IEEE International Conference on Power Systems*, 2016.

[9] Z. X. J. X. B. Li and Q. H. Wu, "Application of improved gm(1, n) models in annual electricity demand forecasting," in *Innovative Smart Grid Technologies - Asia*, 2016, pp. 1–6.

[10] X. W. Kun Ding, Li Feng, "Forecast of pv power generation based on residual correction of markov chain," in *International Conference on Control, Automation and Information Sciences*, 2015, pp. 355–359.

[11] K. Hafeez and S. Khan, "Risk management analysis with the help of lightning strike mapping around 500 k-v grid station using artificial intelligence technique," in *International Conference on Robotics and Artificial Intelligence*, 2012, pp. 165–168.

[12] W. G.-d. ZHU Xi-an, "Discussion on the standard of non-dimensional method in entropy method," *Statistic and Decision*, no. 2, pp. 12–15, 2015.