Regional analysis on model validation with different metrics

Danqing Wu

The 28th Research Institute of China Electornics Technology Group Corporation

Email: wudanqing86@163.com

Abstract. Compared with the global analysis, the regional analysis can offer more comprehensive and more detail information with an input variable fixed in an interval. Existing validation metrics mostly focus on the global discrepancy between the observations and the computer models. To further measure the difference in detail, the concept of regional validation was proposed in this work. To make regional validation specific and easily understandable, two existing validation metrics—the area metric and the $G$ metric were applied. The contribution to whole validation (CWV) metric was proposed to depict the relationship between the proposed regional validation and the initial whole scale validation. At last, an example was employed to demonstrate the rationality and necessity of the regional validation.

1. Introduction

Limited by the complex engineering environment and the huge cost of physical experiments, the engineers consider computational models evaluated with different methods[1-5] to simulate complex physical phenomena. Model validation is then proposed to measure the discrepancy between the physical experiments and predictions.

Researchers have made key contribution concerning general strategy of conducting model validation. Among the different metrics, the area metric-based methods proposed by Ferson et al.[6] are classical and widely used due to their favorable characteristics. The authors have proposed a new validation metric—the $G$ metric. It has a direct comparison of responses without information loss. In addition, it is intuitionistic and easy to be understood. The classical area metric and the $G$ metric are applied to do regional validation to discuss its properties in this paper.

Regional validation shows validation results in different intervals, however, it does not figure out the relationship between the regional and the whole scale validation. Considering the shortage of regional validation mentioned above, a regional validation measure is presented in this paper, and it can represent the effect of the internal regions of each input variable on the whole model validation. Similar to the definition of contribution to sample mean (CSM) [7], the presented regional validation measure is described by the contribution to whole validation. The CWV can effectively inform the engineers if the data in a given interval can make positive contribution to the whole validation.

The remainder of the paper is organized as follows. In Section 2 the definitions of the area metric and the $G$ metric are reviewed. Regional validation is executed with two validation metrics in Section 3. The concept of CWV is explored and discussed in Section 4. An engineering example is employed to illustrate the rationality of the mentioned methodologies in section 5. Finally, some conclusions are drawn in section 6.
2. Review of the area metric and the G metric

2.1. A brief introduction of the area metric

As mentioned in section 1, various types of validation metrics have been studied by researchers. Among the existing validation metrics, the area metric is widely used due to its excellent features. It can be defined by Equation (1):

\[ d(F^m, S^n_e) = \int_{-\infty}^{\infty} |F^m(y) - S^n_e(y)| \, dy \]  

where \( F^m \) is the prediction cumulative distribution function (CDF) and \( S^n_e \) is the empirical CDF of the observations.

The area metric quantifies overall error and can tell a best choice from several candidate computer models. The area metric shows excellent characteristics for model with single response.

To validate model with multiple correlated responses and avoid the estimation of multivariate empirical CDF of the experimental data, Wei Li proposed the PIT area metric [8]. It uses multivariate probability integral transformation to reduce dimension of original responses. The PIT metric firstly estimates the joint CDF of all model responses, namely \( F^m(y_1,\ldots,y_{d},\ldots,y_i) \). Secondly, the experimental data sets are transformed into a one-dimensional data sequence \( v_j \), in which every \( v \)-value is the joint CDF value of the relevant set of data, that is \( v_j = F^m(y_1,\ldots,y_{d},\ldots,y_i) \). Thirdly, the joint CDF of the responses is transformed into a univariate CDF \( K^m(v) \) based on the multivariate probability integral transformation, where \( v \) is the multivariate PIT random variable of the model responses. The empirical CDF \( S^n_e(v) \) of \( v \) can be estimated as follows:

\[ S^n_e(v) = \frac{\sum_{i=1}^{M} I(v_i, v)}{N} \]  

where \( I(v_i, v) = \begin{cases} 1 & v_i \leq v \\ 0 & v_i > v \end{cases} \) and \( N \) is the number of the observations.

The PIT metric is then measured by the area difference of \( K^m(v) \) and \( S^n_e(v) \), which can be described as:

\[ d(K^m, S^n_e) = \int_{0}^{1} |K^m(v) - S^n_e(v)| \, dv \]  

From the definition of the area metric and the PIT area metric, both metrics need to estimate the joint CDF of the computer models, and the former also need compute the empirical distribution of the physical observations. In this work, when the output is unidimensional, the initial area metric is applied; and when output is multidimensional, the PIT area metric is used. Thus, to make the expression compact, these two metrics are both expressed as the area metric.

2.2. A review on the G metric

The G metric is based on the idea of model-free sampling. Model-free sampling can be simply described as a process to extend a sample with small number of elements into that with a larger number of elements [9].

The original observed sample with a small number of elements is denoted as follows:

\[ S_0 = \{s_{0,i} \} \quad (i = 1,\ldots,m_0) \]  

where \( m_0 \) is the number of observations.

The extended sample with a larger number of elements is described as:

\[ S_1 = \{s_{1,j} \} \quad (j = 1,\ldots,m_1) \]  

where \( m_1 \) is the number of extended sample and \( m_1 > m_0 \).
To insure that the extended sample is statistically consistent with the original sample, two criteria in model-free sampling: the assignment criterion and the distance criterion, have been proposed for sample point with a single coordinate in Reference [9]. When each sample point has more than one coordinate, the criterions can be described in Equation (6) and Equation (7).

The assignment criterion aims to assure that each element in 0\(S\) is assigned same number of elements from 1\(S\). The assignments of elements in 1\(S\) depend on the Euclidean distance between 1\(s_{ij}\) and 0\(i_s\). Every element in 1\(S\) is assigned to 0\(s_{ij}\) with the shortest Euclidean distance. \(n(s_{0j})\) is assumed as the number of the assigned elements to 0\(s_{ij}\), and the target value of \(n(s_{0j})\) is \(m_i / m_0\). As a result, the assignment criterion is the minimum of \(C_1\):

\[
C_1 = \sum_{i=1}^{m_0} \left( n(s_{0j}) - \frac{m_i}{m_0} \right)^2
\]  

(6)

The distance criterion aims to reflect the consistency of elements’ locations in 0\(S\) and 1\(S\). The target value of the distance criterion is the minimum of \(C_2\):

\[
C_2 = \sum_{j=1}^{m_1} d[s_{0j}(s_{1j}), s_{ij}]^2
\]  

(7)

where \(s_{0j}(s_{1j})\) shows that \(s_{0j}\) is the element which is the nearest one to \(s_{1j}\). \(d[s_{0j}(s_{1j}), s_{ij}]^2\) is square of the Euclidean distance between \(s_{0j}(s_{1j})\) and \(s_{ij}\).

Beer constructed a dissimilarity measure \(G\) for real-valued samples 1\(S\) and 0\(S\), the assignment criterion according to Equation (6) and the distance criterion according to Equation (7) are combined. As a standard formulation, the quantity is selected as follows [8]:

\[
G = \sqrt{C_1 + C_2}
\]  

(8)

The smaller the \(G\)’s value is, the better statistical consistency shows between 1\(S\) and 0\(S\). The minimum value of \(G\) indicates that elements in 1\(S\) symmetrically and serriedly distribute around each element in 0\(S\). With data in 0\(S\) refers to the physical observations and 1\(S\) refers to the computer model, \(G\) in Equation(8) can be used to distinguish the difference between the physical observations and the computer models. For model with multiple responses, the \(G\) metric can be written as follows:

\[
G = \sqrt{\sum_{i=1}^{M} \left( n(s_{0i}) - \frac{M}{N} \right)^2 + \sum_{j=1}^{N} d[s_{0j}(s_{1j}), s_{ij}]^2}
\]  

(9)

where \(N\) is the number of the observations and \(M\) is the number of computer data.

Apparently, the increase of the dimension of responses does not cause much computation in metric \(G\). Furthermore, the consideration of epistemic uncertainty will enhance the complexity of estimating responses but not increase difficulty in the computation of \(G\) metric.

According to its definition, \(G\) is a colligation of the discrepancy between every physical observation and its corresponding data in computer models. In another word, \(G\) describes the discrepancy starting from a detail difference and ending with a whole difference. Generally, \(G\) metric can distinguish different computer models through its different values. A smaller value suggests a better computer model, and inversely, a bigger value indicates a worse one.

3. Regional validation with two validation metrics

Considering the characteristic mentioned above, if the area metric and the \(G\) metric are applied for model validation in a reduced interval, the results can reflect the discrepancy over the interval.

The variables in a model can be sorted in three kinds: input variables \(x\), model parameters \(\theta\) and site variables \(z\). \(x\) and \(\theta\) can be either deterministic or undeterministic. \(z\) are deterministic space or time domain variables representing the sites to validate the model. The observations can be described
as \( y^e = f(x,z) \), where \( x \) are undeterministic and \( z \) are deterministic. Considering a mathematical or computational model with the form \( y^m = f(x,z,\theta) \), where the model parameters \( \theta \) are deterministic. \( x = (x_1,x_2,...,x_n) \) is the \( n \)-dimensional independent input variables with uncertainty.

The definition of regional validation (RV) for a given input \( x_i \) is as follows:

\[
RV_{x_i}(q) = \text{metric} \ (x_i \in [F_{i}^{-1}(0),F_{i}^{-1}(q)])
\]  

(10)

where quantile \( q \in (0,1] \), \( F_{i}^{-1}(0) \) and \( F_{i}^{-1}(q) \) are separately the inverse CDF of \( x_i \) at quantile 0 and \( q \).

Equation (10) can have different forms according to different validation metrics. The corresponding regional validation metrics of the area metric and the \( G \) metric can be expressed as follows:

The area metric:

\[
\text{Single Response: } S_{x_i}(q) = \int_\Omega F^m(y_{x_i,q}) - S_N^e(y_{x_i,q}) \, dy_{x_i,q}
\]  

(11)

\[
\text{Multiple Responses: } S_{x_i}(q) = \int\int F^m(v_{x_i,q}) - S_N^e(v_{x_i,q}) \, dv_{x_i,q}
\]  

(12)

where \( y_{x_i,q} \) is the corresponding response when \( x_i \) is fixed in the interval \([F_{i}^{-1}(0),F_{i}^{-1}(q)]\), and \( v_{x_i,q} \) is the joint CDF value of the corresponding responses when \( x_i \) is fixed in the interval \([F_{i}^{-1}(0),F_{i}^{-1}(q)]\), that is \( v_{x_i,q} = F^m_{x_i,q}(y_{x_i,q}) \), \( y_{x_i,q}, F^m_{x_i,q}(v_{x_i,q}), S_N^e(v_{x_i,q}) \) are the same as those in the whole scale validation.

The \( G \) metric can be written as follows in the case of either single response or multiple responses:

\[
G_{x_i}(q) = \sqrt{C_1(q) + C_2(q)}
\]  

(13)

where \( n_{x_i,q} \) is assumed as the number of the assigned elements to \( s_{ij} \) in the chosen observations when \( x_i \) is fixed in the interval \([F_{i}^{-1}(0),F_{i}^{-1}(q)]\), and \( d_{x_i,q} \) is the Euclidean distance between \( s_{ij} \) and \( s_{ij} \). \( s_{ij} \) and \( s_{ij} \) have the same meanings as those in the whole scale validation, only differ in the interval of the observations and the computer data.

Compared with model validation in the whole scale of the inputs, regional validation can show detail difference in the arbitrary region \([0, q]\). As a result, regional validation can reflect much more information than traditional validation. When \( q = 1 \), the interval validation is just the whole scale validation. Apparently, the whole-scale validation is a special case included by regional validation, so the latter is more comprehensive and effective for engineers.

4. A development of the regional validation—the CWV metric

As mentioned above, the difference in any interval can be detailedly shown by studying regional validation. However, how a result of regional validation infects the whole validation result is an unknown issue. To find out the answer, a new proportional metric—the CWV (contribution to the whole validation) metric is proposed here. Considering the application of two validation metrics in this paper, the CWV can be in the form of the area metric (SCWV) or in the form of the \( G \) metric (GCWV).

\[
\text{SCWV}_{x_i}(q) = \frac{S_{x_i}(q)}{S_w}
\]  

(14)

\[
\text{GCWV}_{x_i}(q) = \frac{G_{x_i}(q)}{G_w}
\]  

(15)
where the numerator $S_i(q)$ or $G_i(q)$ refers to validation result in the interval $(0, q)$ of $x_i$. The denominator $S_w$ or $G_w$ refers to validation result in the whole scale. Apparently, Equation (18) and (19) offer the engineers another visual perspective to comprehend the regional validation. The different values of the ratio represent different information. If the ratio $SCWV_i(q)$ or $GCWV_i(q)$ is larger than 1, it means that the data in the interval have a negative effect on the whole validation, in another word, the data have a bad match with the observations in the interval and thus make the value of $S_w$ or $G_w$ has a trend of getting larger. On the other hand, if the ratio is smaller than 1, it means that the data in the interval have a positive effect on the whole validation. In general, the CWV metric answers how the difference between the observations and the computer models in a given interval affects the whole difference between them. As a result, the CWV metric can check out if the used validation metrics (the area metric or the $G$ metric in this paper, and also can be other validation metrics) are stable with the increase of computer data.

5. An engineering example
To further demonstrate the performance of the proposed metrics to engineering applications, we consider an automobile front axle in this work. In the automobile engineering, the front axle beam is used to carry the weight of the front part of the vehicle (See Figure 1). As the complete front part of the body rests on the front axle beam, it must be robust in construction. Nowadays, the I-beam structure enjoys the popularity in the design of front axle due to its high bend strength and light weight. The maximum normal stress and shear stress of the I-beam part are $\sigma = M / W$ and $\tau = T / W_\rho$ respectively, where $M$ and $T$ are bending moment and torque, $W$ and $W_\rho$ are section factor and polar section factor. The $\sigma$ and $\tau$ can be expressed as:

$$\sigma = \frac{M}{6h} \left[ \frac{\theta_1(h-2z)^3}{6h} + \frac{\theta_2}{6h} (h^3 - (h-2z)^3) \right]$$

$$\tau = \frac{T}{0.8 \theta_e z^2 + 0.4 \theta_e (h-2z)/z}$$

where the validation site $z = 14$, model parameters $\theta_1 = 12$ and $\theta_2 = 65$. The variables of $h$, $M$ and $T$ are independent normal variables with distribution parameters listed in Table 1.

![Figure 1. The schematic diagram of automobile front axle.](image)

| Random Variables / Random Variables | $h$ / mm | $M$ / (N·mm) | $T$ / (N·mm) |
|-----------------------------------|----------|--------------|--------------|
| Mean $\mu$                       | 85       | $3.5 \times 10^6$ | $3.1 \times 10^6$ |
| Standard deviation $\sigma$      | 0.425    | $1.75 \times 10^3$ | $1.55 \times 10^3$ |

Three candidate computer models are described as follows:
mod el 1 \[
\begin{cases}
\sigma_{1}^{m} = \sigma^{e} \\
\tau_{1}^{m} = \tau^{e}
\end{cases}
\quad (\theta_1 = 12, \theta_2 = 65)
\]

mod el 2 \[
\begin{cases}
\sigma_{2}^{m} = \sigma^{e} \\
\tau_{2}^{m} = \tau^{e}
\end{cases}
\quad (\theta_1 = 10, \theta_2 = 65)
\]

mod el 3 \[
\begin{cases}
\sigma_{3}^{m} = \sigma^{e} \\
\tau_{3}^{m} = \tau^{e}
\end{cases}
\quad (\theta_1 = 10, \theta_2 = 63)
\]

Table 2 describes the results of the whole scale validation. The results of the regional validation of the area metric and the $G$ metric are separately shown in Figure 2 and Figure 3.

| The area metric $S_w$ | The $G$ metric $G_w$ |
|----------------------|---------------------|
| Model 1 | 0.0164 | 212.3 |
| Model 2 | 0.1661 | 574.9 |
| Model 3 | 0.2510 | 1339.5 |

The data in Table 2 illustrate that both the area metric and the $G$ metric can distinguish the different computer models in the whole scale. In brief, model 1 is the best model in describing the physical experiments, followed by model 2, and model 3 is the worst one.

According to Figure 2 and Figure 3, some conclusions can be drawn: Firstly, model 1 shows absolute ascendancy compared with model 2 and model 3 for either the area metric or the $G$ metric in regional validation. In other word, model 1 is close to the physical experiments both in detail and in the whole scale. Secondly, for the area metric, when $h$ is fixed or $M$ is fixed, the regional validation results show a slight concussion around the whole scale validation result, i.e. $S_w$. When $T$ is fixed, the regional validation results show a slowly decrease and concussion with the increase of $q$. It means
compared with $h$ and $M$, $T$ is more sensitive to the change of the interval. Thirdly, for the $G$ metric, with the increase of $q$, the discrepancy of the models increase, which means that with more observations usable, it is more easier to distinguish different computer models.

The CWV metric of two validation metrics are described separately in Figure 4 and Figure 5.

Figure 4. The SCWV metric when (a) $h$ is fixed; (b) $M$ is fixed; (c) $T$ is fixed.

Figure 5. The GCWV metric when (a) $h$ is fixed; (b) $M$ is fixed; (c) $T$ is fixed.

Figure 4 and Figure 5 indicate that: Firstly, for the SCWV metric, its values of model 2 and model 3 are larger than 1, and it indicates that the increase of data has positive influence on the whole scale validation. Because of its smallest value of $S_w$, mode 1's SCWV shows a concussive variety even with a slight change. Secondly, for the GCWV metric, the conclusions are similar. Figure 5 (a) and Figure 5 (b) describe that three models are in increase tendency with the increase of $q$. When $T$ is fixed, only model 1 stays a windless increase, however both model 2 and model 3 show huge changes. Thirdly, considering the SCWV metric and the GCWV metric for the right model (model 1), the validation metric of latter, i.e. the $G$ metric, shows a more stable feature when the interval changes.

6. Conclusions

The existing model validation metrics mostly focus the global difference and ignore the detail difference between the observations and the computer data, while the latter may offer key information for model validation and calibration. Regional validation was proposed to check the detail information of model validation. To make the concept intuitionistic, the area metric and the $G$ metric were applied as comparison. The CWV metric was proposed to describe the relationship between the regional validation and the whole scale validation. Regional validation measures detail discrepancy in different intervals and the CWV metric depicts the influence the regional validation has on the whole scale validation. Generally, the whole scale validation judges the good or bad computer models, while the regional validation describes the partial discrepancy and variety tendency of the models.

References

[1] Al-Musaylh M S, Deo R C, Li Y and Adamowski J F 2018 Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode
decomposition with adaptive noise for multiple-horizon electricity demand forecasting.

*Applied Energy* 217 pp. 422-439

[2] Deo R C and Şahin M 2017 Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland *Renewable and Sustainable Energy Reviews* 72 pp. 828-848

[3] Wang D, Luo H, Grunder O, Lin Y and Guo H 2017 Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm *Applied Energy* 190 pp. 390-407

[4] Al-Musaylh M S, Deo R C, Adamowski J F and Li Y 2018 Short-term electricity demand forecasting with MARS SVR and ARIMA models using aggregated demand data in Queensland, Australia *Advanced Engineering Informatics* 35 pp. 1-16

[5] Prasad R, Deo R C, Li Y and Maraseni T 2017 Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm *Atmospheric Research* 197 pp. 42-63

[6] Ferson S, Oberkampf W L and Ginzburg L 2008 Model validation and predictive capability for the thermal challenge problem *Computer Methods in Applied Mechanics and Engineering* 197 2408-2430

[7] Bolado L R, Castaings W and Tarantola S 2009 Contribution to the sample meanplot for graphical and numerical sensitivity analysis *Reliab. Eng. Syst. Saf.* 94 1041-1049

[8] Li W, Chen W, Jiang Z, Lu Z Z and Liu Y 2014 New validation metrics for models with multiple correlated responses *Reliability Engineering and System Safety* 127 1-11

[9] Beer M 2007 Model-free sampling *Structural Safety* 29 49-65