Modelling error evaluation of ground observed vegetation parameters

Boyi Liang 1,2, Cecilia A. L. Dahlsjö 2, Victoria Maguire-Rajpaul 2, Yadvinder Malhi 2 and Suhong Liu 3,*

1 College of Urban and Environmental Sciences, Peking University, Beijing, China 100871; liangboyi@pku.edu.cn
2 Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom OX1 3QY; c.dahlsjo@gmail.com (C.A.L.D.); victoria.maguirerajpaul@wolfson.ox.ac.uk (V.M.-R.); yadvinder.malhi@ouce.ox.ac.uk (Y.M.)
3 Faculty of Geography, Beijing Normal University, Beijing, China 100875
* Correspondence: liush@bnu.edu.cn

Abstract: To verify large-scale vegetation parameter measurements the average value of sampling points from small-scale data are typically used. However, this method undermines the validity of the data due to the difference in scale or an inappropriate number of sampling points. A robust universal error assessment method for measuring ground vegetation parameters is therefore needed. Herein, we simulated vegetation scenarios and measurements by employing a normal distribution function and the Lindbergh-Levi theorem to deduce the characteristics of the error distribution. We found that the small- and large-scale error variation was similar among the theoretically deduced Leaf Area Index (LAI) measurements. Additionally, LAI was consistently normally distributed regardless of which systematic error or accidental error was applied. The difference between observed and theoretical errors was highest in the low-density scenario (7.6% at <3% interval) and was lowest in the high-density scenario (5.5% at <3% interval) while the average ratio between deviation and theoretical error of each scenario was 2.64% (low-density), 2.07% (medium-density) and 2.29% (high-density). Further, the relative difference between theoretical and empirical error was highest in the high-density scenario (20.0% at <1% interval) and lowest in the low-density scenario (14.9% at <1% interval), respectively. These data show the strength of a universal error assessment method and we recommend that existing large-scale data of the study region are used to build a theoretical error distribution. Such prior work in conjunction with the models outlined in this paper could reduce measurement costs and improve the efficiency of conducting ground measurements.

Keywords: error assessment; vegetation parameters; LAI; ground observation; measuring method

1. Introduction

Establishing the connection between genotype and phenotype is currently one of the most significant challenges facing modern plant biology [1]. Measurements of different vegetation parameters can help us understand genetic characteristics [2]. The utility and importance of terrestrial vegetation (including crops) parameters, such as Leaf Area Index (LAI) and Fractional Vegetation Cover (FVC), have increased in recent years [3-6]. There are two universally recognised methods for measuring these parameters including (a) remote sensing inversions [7-9] and (b) ground observations [10, 11]. Remote sensing inversion directly measures vegetation parameters at large scales (few meters to hundreds of kilometres). However, due to the technical limitations of remote sensing (relatively narrow spatial and temporal resolution, and uncertainty of methodology) it is often necessary to cross-validate these data with ground observations [12-14]. Meanwhile, as the extensive collection of phenotypic data remains onerous, there is often a focus on traits that are easy or inexpensive to measure, while more costly or difficult-to-score phenotypes are studied in only a few individuals [15, 16]. This approach is bound to create uncertainty when it comes to generating the true value of each vegetation parameter. In contrast to remote sensing, no universal methods for
error assessments of ground observations exist. Survey costs and land accessibility limit the extent to which ground observations can be measured, and so data are normally extrapolated based on parameters calculated for small areas [17-19]. For these reasons, whole regions are rarely or never measured in their entirety which means that sampling errors always exist.

Vegetation parameters contain two main error components [20] including a systematic error that varies between different instruments’ attributes or protocols, and accidental error made by the surveyors [21]. While probability theory, such as likelihood theory [22] and Bayes theory [23], have been used to improve existing models, the chosen error assessment must be based on appropriate specificities for each model [24]. Analytical precision can be measured by analyzing replicates or in combination with sampling precision using a balanced design of sampling and analytical duplicates. However, there are no general methods for estimating sampling bias [25]. Besides, all the previous methods for error assessment were conducted after the measurement of vegetation parameters and we had no expectation of error distribution prior to field work [26, 27]. To address these issues, we took LAI as representative of vegetation parameters, and the aims of this study are to (a) create an equation that can be used to estimate error distribution for ground observations, and (b) test the deduced equation on a virtual scenario using ground observations of LAI and compare empirical and theoretical error distributions.

2. Deduction of ground-based error assessment

In this section, we go through the process of deduction of error distribution from normal distribution theorem and Lindeberg-Levi theorem (the latter is also known as independent distribution centre limit theorem). To our knowledge, this is the first time that LAI measurements have been deconstructed (true LAI, systematic error, and accidental error) and combined with the Lindeberg-Levi equation to achieve theoretical distribution values. Using this method, we enable error distributions to be calculated which are important for evaluating accuracy of vegetation parameter measurements.

The Lindeberg-Levi theorem states that when a sampling method for an independent variable (mean $\mu$, standard deviation $\sigma$) is used, the mean tends to be normally distributed as long as the sampling size ($n$) is adequately large [28, 29]. This is expressed as:

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \sim N(\mu, \frac{\sigma^2}{n})$$ (1)

Where $\xi_i$ is each measurement, $n$ is the number of total sampling points and $N$ is normal distribution.

The LAI for each sampling point is expressed as:

$$\bar{X} = x + m + \varepsilon$$ (2)

Where $X$ is the LAI measurement for each sampling point, $x$ is the true value of each data point (without any errors), $m$ is the systematic error and $\varepsilon$ is the accidental error for each sampling point (irrespective of gross error). The presence of the two errors leads to a certain degree of deviation from the value, $x$. When ground measurements of LAI are sampled over large areas, it is likely that a variety of instruments and methods are used which may induce systematic errors. We may calculate the average value of the LAI measurements with the following equation:

$$\overline{X} = \bar{X} + \frac{\sum_{i=1}^{q} m_i \times f_i}{\sum_{i=1}^{q} f_i} + \varepsilon$$ (3)

Where $q$ is the different types of systematic errors, $f_i$ is the number of sampling points under each systematic error and $m_i$ is the value of each systematic error.
According to the statistical principle of accidental error, the mean of errors would converge to zero as sample size increases [30]. Thus, regardless of vegetation attributes, equation (3) may be re-written as follows:

\[
\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} X_{ij} - \sum_{j=1}^{p} m_{j} \times f_{j}}{\sum_{j=1}^{p} f_{j}}
\]  

(4)

Where \( p \) is the number of sampling points for each systematic error. Equation (4) suggests that the difference between the true value \( x \) and the actual measured value \( X \) is equal to a weighted average of the systematic errors (accidental error \( \varepsilon \) is eliminated during the averaging of the values).

Based on the above equations we use normal distribution theorem and Lindeberg-Levi theorem to deduce the following:

\[
\frac{1}{n} \sum_{j=1}^{p} \varepsilon_{j} \sim N(\mu - \frac{\sum_{j=1}^{p} m_{j} \times f_{j}}{\sum_{j=1}^{p} f_{j}} , \frac{\sigma^2}{n})
\]  

(5)

Where the result will follow normal distribution based on the total sampling number (n). The mean is the true LAI minus the weighted average of the systematic errors and the variation is \( \frac{\sigma^2}{n} \). As this deduction was based on vegetation density and research area, the distribution and forest type will not have an impact on the results. Rather it should be used as an error assessment tool. Equation (5) should be used to assess the distribution of the sampling error as well as the probability of different error intervals (deviation from true LAI) based on the variety of systematic errors and variation of LAI across the whole study region. If the variation of LAI at different scales are similar, we assume that the variation is the same for the whole region. In the simulated model below, we test the existence of such an assumption.

### 3. Theoretical Validation of Error Assessment

To validate equation (5), we used Matlab software to create three scenarios with low-, medium-, and high-density vegetation levels (representing LAI of 2.54, 5.09 and 8.09, respectively). For building a single tree, three elements including trunk position, trunk height and branch length are imported. In previous studies, researchers found that the position and height of trees in homogeneous forests tend to follow normal and Poisson distribution respectively [31-32]. Based on these assumptions, we created two groups of random numbers representing tree height and tree position. Then for each group, three numbers were randomly selected to create a low-density scenario. Meanwhile, five and eight numbers were selected in order to create a medium-density and high-density scenario respectively. Branches (left/right) were created using equation (6) and (7) where the tree trunk was set perpendicular to the ground at an angle of \( \alpha = \pi / 2 \) and the relationship between the angles of the upper branches (\( \alpha_{n+1} \)) and the lower branches (\( \alpha_{n} \)) are described as follows:

\[
\alpha_{n+1, L} = \alpha_{n} + \pi / 8 \\
\alpha_{n+1, R} = \alpha_{n} - \pi / 8
\]

(6)

Where \( L \) is left, and \( R \) is right. The relationship of the length of the upper branches (LOB\(_{n+1}\)) and lower branches (LOB\(_{n}\)) is described as follows:

\[
\text{LOB}_{n+1} = 4 \times \text{LOB}_{n} / 5
\]

(7)
Relevant parameters [33] were used to create three scenarios (low-density level, medium-density level and high-density level) using Matlab, and each scenario includes pixels (Fig. 1) of either vegetation or none-vegetation. The true LAI for each scenario was calculated as ratio of number of vegetation pixels and number of horizontal pixels in breadth.

![Vegetation scenarios](image)

**Figure 1.** Vegetation scenarios (a. low-density level; b. medium-density level; c. high-density level).

In each scenario, the observed value of LAI was calculated by randomly selecting vertical lines where the number of vegetation pixels that each of them contained was summed and divided by the number of lines. Computer simulations were used to calculate the error distribution of different error intervals to verify the authenticity of the theoretical deduction.

![Measurement variances](image)

**Figure 2.** Measurement variances (a) and their anomaly (b) with average value at different sampling scale. There are 512 horizontal pixels in each scenario, where LAI was measured at different sampling scales (1, 2, 4, 8, 16, 32, 64 or 128 pixels, the sample scale choice was denoted as Δx).

The variation of the observed LAI was stable at certain pixel scales with lowest variability across sampling sizes at the medium scales (Fig. 2) suggesting that our assumption (small- and large-scale error variation is similar) was true. Based on the estimated LAI variation at the whole study region, equation (5) was used to calculate the error distribution of LAI measurements.
Table 1. Systematic error setting (TS: Types of Systematic error; Se: Systematic error; Sn: Sampling size). The system error and the number of sampling points are set artificially for different types of systematic errors. In each kind of systematic error, program also simulated accidental error following normal distribution. Theoretical errors were calculated by Equation (4) and (5).

<table>
<thead>
<tr>
<th>TS</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Fourth</th>
<th>Fifth</th>
<th>Theoretical error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se</td>
<td>Sn</td>
<td>Se</td>
<td>Sn</td>
<td>Se</td>
<td>Sn</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>50</td>
<td>0.4</td>
<td>70</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>50</td>
<td>0.4</td>
<td>70</td>
<td>0.6</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>50</td>
<td>0.4</td>
<td>70</td>
<td>0.6</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>50</td>
<td>0.4</td>
<td>70</td>
<td>0.6</td>
<td>90</td>
</tr>
</tbody>
</table>

![Graph](image)
Figure 3. Normal distribution test of measurement results under various systematic errors (a: low-density level; b: medium-density level; c: high-density level). The vertical dashed line represents the mean value of LAI and the slanted dashed line stands for strictly normal distribution.)
Figure 4. Measurement error distribution (a: low-density level; b: medium-density level; c: high-density level, systematic errors (SEs)). Each line represents the measurement scenarios with different types of systematic error (j = 1, 2, 3, 4, 5, as defined in equation (4)), and the abscissa stands for different ratio between deviation and theoretical LAI. For example, <1% means measurement result error is less than 1%. The y axis is the percent (*100) of sampling points in one error interval to total points. The grey area indicates the error of empirical error and the theoretical error (each dashed line). Every measurement scenario is simulated 1000 times.

Error distribution of LAI was calculated using a sequence of normal distribution to represent accidental error as well as five types of systematic error with the number of systematic errors ascending from 50 to 130 (Table 1). Using equation (5) we found that the measured average LAI was consistently normally distributed regardless of which systematic error or accidental error was applied (Fig. 3). Additionally, the differences between the observed and theoretical errors were highest in the low-density scenario (7.6% difference when error interval was <3%) and lowest in the high-density scenario (5.5% difference when error interval was <3%) (Fig. 4). Average ratio between deviation and theoretical error of each scenario was 2.64% (low-density), 2.07% (medium-density) and 2.29% (high-density). The average percentage of empirical error located in one interval is therefore in the region of about 2% of the theoretical value.
Figure 5. Ratio between deviation and theoretical error in different intervals with different types of systematic error (a: low-density level; b: medium-density level; c: high-density level).

Further, the relative difference between theoretical and empirical error (when error interval was <1%) was highest in the high-density scenario (20.1% with four kinds of systematic error) and lowest in the low-density scenario (0.17% with one kind of systematic error). The average deviation from the mean between the theoretical and empirical errors in each scenario was 5.9% (low-density), 4.1% (medium-density) and 4.9% (high-density).

4. Discussion

In the field, different research plots often have different vegetation conditions (such as mean LAI and deviation of LAI in different subplots). However, it is hard to obtain real systematic error of measurements in each location. This is because systematic error is not only caused by different instruments, they are also influenced by the measuring behaviour of each surveyor. For this reason, we used real data to validate error distributions that may have been created during fieldwork using different sampling sizes, excluding systematic error.

The data were gathered from an old secondary forest (BOB-03; 50 years post-logging, 6.69201N, -1.307755W) and a Savannah-forest transition plot (KOG-05, 7.30115N, -1.164933W) in Ghana and from an old growth forest (NXV-02, -14.7075S, -52.3517W) in Brazil. LAI was measured by taking three photos (exposures -1, 0, +1) at 1m height with a fish-eye lens in each sub-plot (25 in each plot). The photos were analyzed with the software Hemisfer 2.2 (see the supplementary materials for details of the settings). Each plot contains 25 subplots and standard deviation of LAI for each plot is 0.39 (BOB-03), 0.67 (NXV-02) and 4.87 (KOG-05) respectively.

We can see in Figure 6 that as sampling quantity increases, the results of measured LAI are closer to mean LAI (we assumed it as true value) of the plot. Meanwhile, for the plot with the highest deviation of LAI, a higher number of samples are needed in order to reduce the deviation. For example, three sampling points in BOB-03 provided a result close to mean value while the deviation of 13 sampling points in KOG-05 was still high (about 1). The LAI in BOB-03 converged with the mean faster than in NXV-02 and KOG-05. In Figure 6b, the results of measurements in BOB-03 is more concentrated around the mean LAI no matter how many sampling points we used. This suggests that the sampling quantity of measurement in one research plot to reach specific error requirement should be decided by deviation of different subplot rather than average LAI of the whole plot. This is particularly important when working with different habitat types as the forest plot in BOB-03 which has a more homogenous canopy.
Errors occur in all geographical measurements regardless of what instruments and methods are implemented [34, 35]. Additionally, both the number of sampling points and the sampling location will have an impact on the measurement output. Data collected at the small scale are often used to verify large scale data or examine relationship between the statistical precision of sample estimates and plot size [36-38]. While these methods have been used as error assessments, it is now clear that measurements at different scales are likely to undermine the validity of data. For example, based on Equation (4) and (5), when the true value of LAI in a study area is 8.09, there is a 50% chance that the deviation will be greater than 5% if the sampling size is 50. However, if the sampling size is 450, there is a high chance that the probability will be less than 5%. Therefore, better understanding and control of the error during sampling will benefit analysis of relationships between measured value and true value of different vegetation parameters.

5. Conclusion

In this study we have demonstrated the reliability and applicability of error assessment in LAI ground observations where any deviation of error distribution could be due to either the number of sampling points or the process of averaging variation across different scales. What is more, this method puts the systematic error and accidental error into the evaluation system, making the results more reliable.

To deal with errors that occur in the field, we should not only focus on promoting the accuracy of instruments [39, 40] and improve the authenticity of models [41, 42], but also pay attention to the distribution of errors. Overall, our error assessment method has two advantages over other models: (a) error assessment is conducted before measurement which will give surveyors an expectation of error distribution, (b) the method has its adaptability and flexibility as theoretical error distribution for each study area is calculated by its vegetation distribution.

For future ground observed measurements we recommend that the average variation of LAI at the large scale (e.g. using drone or satellite imagery) is calculated or that the variation of previous vegetation parameter observations in the area are acquired. These data may thereafter be used to build a theoretical error distribution. Prior knowledge is key to produce more accurate estimates and researchers are encouraged to have an appropriate number of sampling points to reasonably meet the error requirement. Such prior work in conjunction with the models outlined in this paper could reduce measurement costs and improve the efficiency of conducting ground measurements.

Funding: This study was funded by National Key R&D Program of China (2017YFE0118100), National Natural Science Foundation of China (No.41171262 and No. 41861053).

References


