Biomass and Bioenergy 115 (2018) 27–34

Contents lists available at ScienceDirect

Biomass and Bioenergy

journal homepage: www.elsevier.com/locate/biombioe

Research paper

Dealing with small sets of laboratory test replicates for Improved Cooking Stoves (ICSs): Insights for a robust statistical analysis of results

Francesco Lombardi*, Fabio Riva, Emanuela Colombo

Politecnico di Milano, Department of Energy, Via Lambruschini 4, Milan, Italy

ARTICLE INFO

Keywords:
Cookstove
Statistic
Test
Protocol
Performance
ICS

ABSTRACT

Improved Cooking Stoves (ICSs) represent the most commonly promoted solution to alleviate the burden associated with the use of traditional biomass in a short-term perspective. However, criticism is raising about the methodologies used for assessing their performance, with a particular focus on laboratory-based testing protocols. One of the key weaknesses of current protocols consists in the inaccurate and biased approach adopted for reporting and statistically analysing test results, which can lead to misleading conclusions about the actual improvements ensured by ICSs. This study proposes a robust procedure to deal with the statistical analysis of small sample sizes, and subsequently verify it through its application to an experimental comparison – based on the Water Boiling Test – between three models of stove. The results show that the current practice based on 3 or 5 replicates often produces biases in the analyses, as at least 13 replicates might be needed to achieve reliable results. Moreover, the study shows how the t-test is in most cases improperly applied, while the proposed procedure allows to deal both with normally and of non-normally distributed data sets in a robust way. In one case, the apparent improvement of an ICS model as compared to the three-stone fire, is refuted by the application of our procedure.

1. Introduction

Biomass-fuelled Improved Cooking Stoves (ICSs) are commonly promoted as a potential interim solution to the lack of access to clean cooking facilities in developing countries, notwithstanding increasing scientific evidences about the limited real-life benefits they are able to bring as compared to the laboratory-based performance [1–6]. In this framework, an accurate assessment of the performance of ICSs represents a critical issue, which is widely debated in the literature with a particular focus on laboratory-based testing protocols, that represent the most widespread methodology for performance assessment [7–16]. Lombardi et al. [17] identified the main issues related to the existing testing protocols, and notably to the most common, *viz.* the Water Boiling Test (WBT), including: (i) the lack of real-life relevance [1,2,7,17–20], (ii) the low repeatability [10,17,21], and (iii) the inaccuracy of methodologies for the statistical analysis of results [9,10,17]. The present study focuses on the latter major issue, which is particularly relevant since biases in statistical inferences can lead to the promotion of non-significantly improved stoves, regardless of the testing protocol employed. Wang et al. [9] and Riva et al. [10] stress this concept, and highlight two major shortcomings that are common to all the statistical approaches of current testing protocols. The first is related to the minimum number of test replicates – typically just 3 – prescribed to evaluate a stove performance and to perform statistical inferences [17]. As a matter of fact, performing a larger set of test replicates allows to achieve a more reliable value of standard deviation, i.e. a value that is representative of the statistical population [9,10], avoiding potential biases in statistical inferences. However, there is a trade-off between the reliability of the standard deviation and the number of test replicates – and thus the time and effort – required. To this end, Wang et al. [9] discourage relying on a three-replicates standard deviation and finally suggest performing at least 5 replicates as a "rule of thumb" to obtain sufficiently reliable results, though such threshold value has never been counter-checked by any other study. The second shortcoming regards the unjustified assumption of the data set being normally distributed. Indeed, the normality condition is an essential formal prerequisite for the application of the *t-test*, which is nonetheless proposed as a method to perform statistical inferences in a large part of the ICSs testing literature regardless of a prior analysis of data distribution [10]. This practice may easily lead to biased inferences, since performance parameters of biomass stoves are likely to experience deviations from normality when a sufficiently large sample size is considered [10,22]. To our knowledge, there are no studies in the literature that discuss how to deal with non-normally distributed data

* Corresponding author.Via Lambruschini 4, 21056 Milan, Italy.
E-mail addresses: francesco.lombardi@polimi.it (F. Lombardi), fabio.riva@polimi.it (F. Riva), emanuela.colombo@polimi.it (E. Colombo).

https://doi.org/10.1016/j.biombioe.2018.04.004
Received 2 November 2017; Received in revised form 20 March 2018; Accepted 12 April 2018

sets in the framework of ICSs performance assessment.

In order to overcome the abovementioned issues, the goal of the present study is to provide a rigorous and practical procedure – verified via its application to a set of experimental tests on two commercial ICSs models and a three-stone fire – to perform a robust statistical analysis of the results of laboratory tests on cooking stoves, consisting of 3 main phases: (i) a practical guide to identify a minimum reliable number of replicates in an experimental campaign, (ii) an accurate and statistically sound method for analysing the statistical distribution and computing the uncertainty, and (iii) a scheme for comparing indicators of performances between two stoves. The order chosen to present the phases of our procedure reflects the chronological order of their practical fulfilment, as shown in Section 4.

2. Procedure for statistical analysis

The proposed procedure builds and improves upon our report titled "*Guidelines for reporting and analysing laboratory test results for biomass cooking stoves*" [23].

2.1. Identifying a minimum reliable number of replicates

Current testing protocols set the minimum number of test replicates required to have a reliable value of standard deviation at 3, while Wang et al. [9] state that at least 5 replicates should be performed, based on their empirical observations with the WBT. Conversely, we propose a practical guide based on an iterative procedure that allows to check, after each additional test replicate performed, if a sufficient level of reliability for the standard deviation has been achieved or not, regardless of the testing protocol employed. A few simple steps shall be followed:

1 Perform at least $n = 5$ test replicates;
2 For each performance indicator of interest
  a Compute the standard deviation;
  b Calculate the percentage change of the standard deviation (S_n), between the *n-th* and *(n − 1)-th* replicates ($\Delta\%_1$), between the *(n − 1)-th* and *(n − 2)-th* replicates ($\Delta\%_2$), and between the *(n − 2)-th* and *(n − 3)-th* replicates ($\Delta\%_3$), as $\Delta\% = \frac{S_n - S_{(n-1)}}{S_{(n-1)}}$;
  c If $\Delta\%_1$, $\Delta\%_2$ and $\Delta\%_3$ are both less than 10% in absolute terms, n is the minimum number of replicates required. Otherwise, add one more test replicate and iterate back from 2.

The rationale behind this practical guide is to check whether the variation of the standard deviation after the addition of a new test replicate is negligible (i.e. less than 10%). Furthermore, since a low variation of the standard deviation can be fortuitously detected just to be refuted once a further replicate is added, our procedure requires that this condition holds true for at least three consecutive replicates ($\Delta\%_1$, $\Delta\%_2$ and $\Delta\%_3$). If this condition is consistently respected, it is reasonable to assume that the value of standard deviation is approximately representative of the statistical population. Nonetheless, the testers shall consider the possibility that significant outliers – due to systematic errors – may arise in the middle of a consistent trend of low percentage change of the standard deviation. In this case, the outliers should be carefully evaluated and eventually discarded. To this regard, the proposed analysis of the percentage change of the standard deviation might be also seen as a support towards an intuitive identification of systemic errors.

Typically, different performance indicators will require a slightly different minimum reliable number of replicates to meet the criterion. Accordingly, the overall minimum number of replicates will correspond to the value for which such criterion is met for all the indicators that the tester needs to measure.

2.2. Analysis of the statistical distribution and the uncertainty

Given a number of n test replicates that is sufficient to obtain a reliable value of standard deviation, the results shall always be averaged and reported as in Equation (1):

$$\overline{X} \pm U_e \text{ (c.l.\%)} \tag{1}$$

Where:

- $\overline{X}$ is the average value of the $[X_i...X_n]$ observations of the selected indicator of performance (e.g. Thermal efficiency η , Specific consumption SC, Time to boil);
- U_e is the expanded uncertainty of the indicator $\overline{X}$, for a selected confidence level ($c.l.\%$), usually 90% or 95%.

In order to compute U_e, two-steps shall be followed:

1 Verify the normality of the data set $[X_i...X_n]$ by means of a Shapiro-Wilk test, which is the most powerful normality test in the conditions of interest [24,25];
2a If the normality hypothesis is not rejected, U_e shall be calculated based on a t-student distribution [26];
2b If the normality hypothesis is rejected, provide the uncertainty based on the *Chebyshev's inequality* – i.e. the most conservative interval. In this case, $U_e = \sqrt{\frac{1}{\alpha}} S_n$ [27], with α equal to 0.10 or 0.05 based on the desired confidence level, and S_n the standard deviation of the data sample.

Relying on Chebyshev's inequality to compute the expanded uncertainty will lead to safer though larger confidence intervals. The final result is considered acceptable if the computed value of U_e is smaller than $\overline{X}$.

2.3. Comparing the performance of two different stove models

In order to compare the indicators of performance between two stove models and to assess the relative improvements, the two respective confidence intervals for a selected parameter shall be compared. If they do not numerically overlap, it is always possible to provide statistically significant conclusions (e.g. one stove performs better than the other). Conversely, if there is overlapping between the two, the tester shall perform a statistical test to draw significant conclusions. Again, two-steps shall be followed for each performance indicator that needs to be compared:

1 Verify the normality of the data sets $[X_i...X_n]_{Stove1}$, $[X_i...X_n]_{Stove2}$related to the selected performance indicator for each stove, by means of a Shapiro-Wilk test [23,24];
2a If the normality hypothesis is not rejected for each of the two data sets, the selected indicator shall be compared by means of a t-test assuming unequal variances [26];
2b If the normality hypothesis is rejected for at least one of the data sets, the tester shall compare the data sets of the selected indicator by means of a non-parametrical test, i.e. the Mann-Whitney rank sum test [28].

The outlined procedure allows to provide the uncertainty and to perform statistical inferences both in case of normally and non-normally distributed data sets, preventing biases such as the use of t-tests for cases in which they are not rigorously valid. It is also worth noting that this procedure keeps valid even when comparing two samples of different sizes, as both the unequal variances t-test and the Mann-Whitney rank sum test can be performed with unequal sample sizes [29].