Energy for Sustainable Development 20 (2014) 21–29

Contents lists available at ScienceDirect

Energy for Sustainable Development

How many replicate tests are needed to test cookstove performance and emissions? — Three is not always adequate

Yungang Wang [a,*], Michael D. Sohn [a], Yilun Wang [b], Kathleen M. Lask [c], Thomas W. Kirchstetter [a,d], Ashok J. Gadgil [a,d]

[a] Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA
[b] ISO Innovative Analytics, San Francisco, CA 94111, USA
[c] University of California, Berkeley, College of Engineering, Applied Science and Technology Program, Berkeley, CA 94720, USA
[d] University of California, Berkeley, Department of Civil and Environmental Engineering, Berkeley, CA 94720, USA

ARTICLE INFO

Available online 13 March 2014

Keywords:
Cookstove
Berkeley–Darfur Stove
Variability
Confidence interval
Kolmogorov–Smirnov test
Bootstrap

ABSTRACT

Almost half of the world's population still cooks on biomass cookstoves of poor efficiency and primitive design, such as three stone fires (TSF). Emissions from biomass cookstoves contribute to adverse health effects and climate change. A number of improved cookstoves with higher energy efficiency and lower emissions have been designed and promoted across the world. During the design development, and for the selection of a stove for dissemination, the stove performance and emissions are commonly evaluated, communicated and compared using the arithmetic average of replicate tests made using a standardized laboratory-based test, commonly the water boiling test (WBT). However, the statistics section of the test protocol contains some debatable concepts and in certain cases, easily misinterpreted recommendations. Also, there is no agreement in the literature on how many replicate tests should be performed to ensure "confidence" in the reported average performance (with three being the most common number of replicates). This matter has not received sufficient attention in the rapidly growing literature on stoves, and yet is crucial for estimating and communicating the performance of a stove, and for comparing the performance between stoves. We illustrate an application using data from a number of replicate tests of performance and emission of the Berkeley–Darfur Stove (BDS) and the TSF under well-controlled laboratory conditions. Here we focus on two as illustrative: time-to-boil and emissions of $PM_{2.5}$ (particulate matter less than or equal to 2.5 μm in diameter). We demonstrate that an interpretation of the results comparing these stoves could be misleading if only a small number of replicates had been conducted. We then describe a practical approach, useful to both stove testers and designers, to assess the number of replicates needed to obtain useful data from previously untested stoves with unknown variability.

Introduction

About half of the world's population uses biomass as fuel for cooking (IEA, 2004). The smoke from biomass cooking fires was recently found to be the largest environmental threat to health in the world, and is associated with 4 million deaths each year (Lim et al., 2012). This exposure has also been linked to adverse respiratory, cardiovascular, neonatal, and cancer outcomes (Smith et al., 2004; Weinhold, 2011). A 2011 World Bank report notes significant contributions of biomass cooking to global climate change (World Bank, 2011). The contribution to climate change from black carbon (BC) emission from biomass cooking is a topic of growing interest, especially in terms of climate forcing and melting of glaciers (Hadley et al., 2010; Ramanathan and Carmichael, 2008). Current biomass stoves lead to a large burden of disease, and contribute to adverse impacts on local and the global environment. Hence there is substantial interest in developing and disseminating fuel-efficient biomass stoves with reduced emissions (e.g. DOE, 2011). Launched in September 2010, the Global Alliance for Clean Cookstoves (GACC) "100 by 20" goal calls for 100 million homes to adopt clean and efficient stoves and fuels by 2020.

The "three-stone fire" (TSF) is a commonly prevailing cooking method for a large fraction of the population at the base of the economic pyramid. In quantifying the performance of an improved stove, the TSF is commonly used as the baseline. This least expensive class of stove is simply an arrangement of three large stones supporting a pot over an open and unvented biomass fire. A TSF is one of the two stoves we analyzed in this study. We also tested the performance and emissions of the Berkeley–Darfur Stove (BDS) as an exemplar of an improved fuel-efficient biomass cookstove. The BDS was developed at Lawrence Berkeley National Laboratory (LBNL) for internally displaced persons in

* Corresponding author. Tel.: +1 510 495 8065.
E-mail address: yungangwang@lbl.gov (Y. Wang).

http://dx.doi.org/10.1016/j.esd.2014.02.002
0973-0826/

Darfur, Sudan (http://cookstoves.lbl.gov/darfur.php). It is an all-metal precision-designed natural-convection stove, with design features co-developed by iterative feedback from Darfuri women cooks. The BDS by design accommodates Darfuri traditional round-bottom cooking pots and cooking techniques (Fig. 1).

A literature survey of recent laboratory cookstove testing in peer-reviewed journal articles shows widely different numbers of replicate tests (Bailis et al., 2007; Jetter and Kariher, 2009; Jetter et al., 2012; MacCarty et al., 2008, 2010; Roden et al., 2009; Smith et al., 2007). The number of replicates reported in these seven studies range from 1 to 23. However, six out of seven studies have reported results with only 3 or fewer replicates. One then can rightly ask: how many replicate tests do I need to test the performance and emissions of the stove? Answering this question is application specific, and requires greater specificity. For example, the question might be better phrased. For a water boiling test (WBT), how many replicates are needed to estimate the average "time to boil" to within 2 min and with 95% confidence? Or how many replicates are needed to confirm, with 95% confidence, that Stove "A" emits less $PM_{2.5}$ than Stove "B"? These questions exemplify perhaps the most frequently asked questions in planning stove experiments and interpreting their results.

There is no single or simple answer to the number of replicates needed to answer the above questions. The answer depends on the experimental design, how many parameters need to be estimated, and the resulting variability in the stove replicates. In this study, we investigate how to answer the above questions using data from the BDS and TSF water boiling experiments. We show how the number of replicates is linked to uncertainty and variability in the experiments and stove performance. We also show how many replicates are likely needed as various practical performance comparisons, such as "Does Stove A perform better than Stove B?" and "What is the uncertainty in the expected performance of Stove A or Stove B?" Finally, we describe a practical approach to design an experiment to test the performance of a previously untested stove.

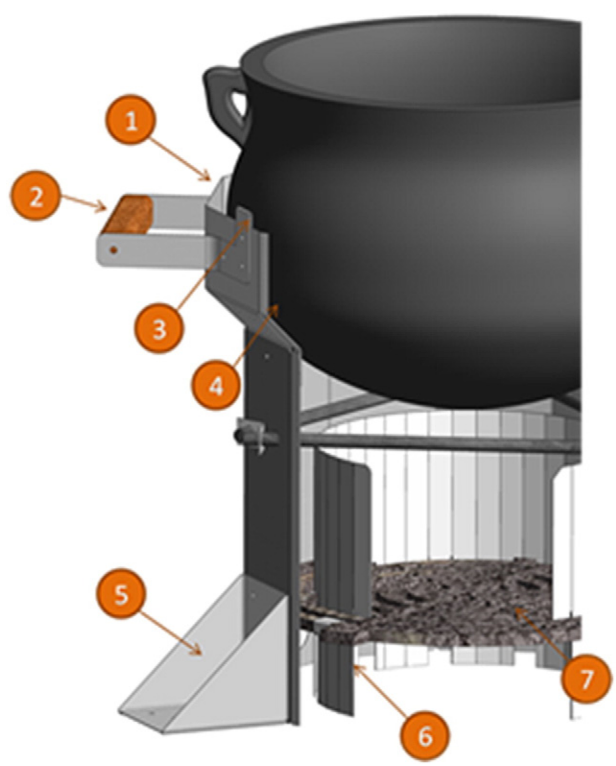

Fig. 1. Schematic of the Berkeley-Darfur Stove. (1) A tapered wind collar that increases fuel-efficiency in the windy Darfur environment and allows for multiple pot sizes; (2) wooden handles for easy handling; (3) metal tabs for accommodating flat plates for bread baking; (4) internal ridges for optimal spacing between the stove and a pot for maximum fuel efficiency; (5) feet for stability with optional stakes for additional stability; (6) nonaligned air openings between the outer stove and inner fire box to accommodate windy conditions; and (7) small fire box opening to prevent using more fuel wood than necessary.

Problem statement and causes of variability

Appendix 6 of the WBT (version 3.0, http://www.pciaonline.org/node/1048) provides a detailed approach for comparing the performance of stoves. It describes a suite of test statistics and important considerations for interpreting test results. While comprehensive, the description contains some debatable concepts and in certain cases, easily-misinterpreted recommendations. For example, it affirms that "At least three tests should be performed on each stove" and provides a cogent explanation for it. It also discusses the importance of paying attention to the statistical significance of a series of comparison tests between the performances of two stoves. While both statements are correct, it is not surprising that stove testers misinterpret these comments as (i) "only three tests are needed" or (ii) a hypothesis test with strong p-value (assuming a Gaussian distribution) provides unarguable confirmation of stove performance or comparison results. In fact, neither interpretation is correct or claimed in the text. We reason further that elucidation of Appendix 6 is necessary, and a more transparent methodology would greatly benefit stove testers. We believe that a transparent methodology would be best accomplished by developing an approach that maps the trade space between sample size, variability, and confidence. We also believe it is important to show that alternative methods for comparing the performances of stoves are available and should be considered. This work thus builds and improves upon Appendix 6 by providing new methods of interpreting test results for stove testers.

The literature generally shows that even under carefully controlled conditions, stove test results show high test-to-test variability (coefficient of variation > 1.0, e.g. Jetter et al., 2012). There are many possible causes of this variability even within a precisely defined test such as the latest WBT (version 4.2.2), and we list a few here. Stove efficiency and emissions are generally a function of thermal power, and owing to the discrete nature of fuel-feeding events, a stove's thermal power invariably varies, also contributing to temporal variability within a test, which can translate into test-to-test variability. Despite due care, the ratio of bark to sapwood to hardwood for various pieces of fuelwood can be different, and thus will have different burn characteristics. Furthermore, different pieces of fuelwood may have different surface to volume ratios, contributing to different rates of burning. Lastly, even reasonably experienced and careful stove testers demonstrate some variability in the way they tend the fire in the stove from test to test, and within a test (Granderson et al., 2009). All these (and other uncontrolled factors) together give rise to what we lump together as variability in the test-to-test replicate results for a stove under controlled laboratory conditions.

Approach

The question of "How many replicate tests do I need?" is not novel. It is a well-researched question in classical statistical theory, but has not received much attention from the stove research community. We briefly summarize here the statistical background relevant to answer the question.

Probability density function and cumulative distribution function

Technically, for a continuous random variable, the probability density function (PDF) describes the probability that a value will be within a certain range of the sample. However, as this range is evaluated by integrating, it can be chosen to be quite small, so for most practical purposes, the PDF may be considered the probability of obtaining a