**Abstract**

Yellowstone National Park (YNP), Wyoming USA, contains over 10,000 geothermal features and 2 to 5% of these features are geysers. Yellowstone has about half of the world’s geysers and the majority of YNP geysers are located in Upper geysers Basin. Beginning in 1970, details (time of eruption, height, duration, etc.) of about 25 geysers activities have been recorded in log books and later transcribed into an electronic dataset and posted on the park’s website. The data was collected by park rangers, visitors, and geyser enthusiasts, among others. The data collected by direct observation, camera, electronic, etc. The dataset contains a great deal of information that is relevant to scientists, educators and the public. However, the use of the dataset is severely limited without cleaning and standardization. Given the size, time span over which the data was collected, and the number of people involved in collecting the data, it’s inevitable that the data contains many inconsistencies. The dataset has been cleaned, standardized, reorganized in some parts and converted to a spreadsheet which makes the dataset much better suited for computations and analysis. The reorganization consists of two steps: step one was to remove text type information and extra information to a newly created column; and step two was to reorder the information in a set of records so that individual data entry is consistent with the column heading under which it should have been listed. The overall and monthly statistical summary of the data shows that interval and duration are both bimodal normally distributed, height is normally distributed and preplay display a Rayleigh type distribution. Comparison of the YNP and the electronic dataset was not feasible for all geysers and all variables; however, where it’s feasible such as the case for interval data, the two datasets are nearly identical.

**Key word:** Datasets- Data cleaning- Data standardization and analysis- Yellowstone National Park- Geysers.

**Introduction**

Yellowstone National Park (YNP), Wyoming, is one of the United States most popular parks. The main attraction of the park is a huge system (hundreds) of geysers; each with variable eruption parameters (e.g., duration, height, etc.). Some of the most popular ones, most of which are fairly close together and easily accessible, were grouped into a “system” and monitored. The best known geyser in the park is the Old Faithful geyser due to its well documented eruption pattern and intensity. The geology of the geyser system in general and Old Faithful in particular is beyond the objective of this paper; they have been studied extensively elsewhere (e.g. Birch and Kennedy, 1972; Dowden, et al. 1991; Hutchinson et al. 1985; Rinehart, 1980; O’Hara and Esawi, 2013).

Park rangers, volunteers, geyser enthusiasts, and at times visitors have been collecting and recoding data about the geyser system’s activities for a long time. The data were collected by several means including direct observation, electronic means, and webcam, among others. Beginning in 1970, the data had been recorded in log books which later were transcribed by Lynn Stephens, Marion Powell and Mary Beth Schwarz of YNP into an electronic dataset (logs) posted at the Geyser Observation and Study Association (GOSA) (<http://gosa.org/ofvclogs.aspx>); without their effort, this dataset would not have been possible. Therefore any suggestion about the quality or shortcomings of the dataset in this paper is only intended for the sole purpose of improving the dataset so that it will be of use to a larger audience. In this study the term dataset refers to the data listed on the GOSA site or any of its modified versions in this paper, unless stated otherwise. The dataset which is organized into several main columns/headings, contains a great deal of information, and covers the period between 1970 and 2011 with a gap w between 1981 and 1991 where there is no data.

In addition to the YNP dataset, there is another dataset that was compiled electronically by Ralph Taylor (<http://www.geyserstudy.org/electronicsummary.aspx>). This dataset contains data for 24 geysers, covers the period from mid-1990s to 2011, and contains interval data for all geysers and duration data for 10 geysers. A recent attempt to consolidate all of the YNP geysers data can be found at GeyserTimes (<http://geysertimes.org/>). This site is very well organized, contains data from both the YNP and the electronic data, contains a glossary for geysers’ terminology, and has more recent data than 2011.

Given the great wealth of information in the dataset and the absence of any data standardization makes the dataset of limited use for basic research and science education. This study thus aims at (a) cleaning and standardizing the dataset and (b) assessing the accuracy and consistency of the dataset. Data cleaning is accomplished by standard data cleaning protocols (e.g. Rahm and Hai Do, 2000; Hellerstein, 2008). Assessing the accuracy and consistency of the dataset is a little more challenging prospect and will be explored by (1) using comparative statistics; (2) comparing and contrasting subsets of the dataset that have been collected via different modes (e.g. visitor(vr), electronic, cam, etc.); and (3) comparing the dataset with the electronically collected data by Ralph Taylor. To achieve a balance between maintaining the integrity of the original data and making it useable for research and education purposes, the number of changes that have been made are kept to a minimum and only changes were made when absolutely necessary; details of the changes made will be described later. Furthermore, over 98% of the changes made are reversible. MS Excel and R (<http://www.r-project.org/>) are the main tools used for data analysis and cleaning. To make the dataset publically accessible to a larger audience, I will explore the possibility of posting it on either the journal’s, YNP, or GeyserTimes websites

Table 1



Table 2

