Data quality checkup

By Jaroslav Mohapl

Abstract

Data quality assurance is of key importance for long term monitoring of the environment. This article suggests a method for improvement of the quality control process and related policies.

Key Phrases Data quality assurance, measurement precision, quality assurance.

Key Words Measurement, precision, data, quality, assurance.

Introduction

This note recalls a simple method for a checkup if the randomness of data sampled by a data collector or data logger is due to natural causes or if a source of a bias is present. In case a bias is detected, and no a reason for that is mentioned in the sampling protocol, the researcher might want to investigate the origin of the data more thoroughly before starting further analysis.

If data such as temperature, wind speed or humidity are presented with precision, say, one decimal, then it is natural to expect each of the decimal digits 0 - 9 to occur with roughly the same frequency. For small data sets we usually do not need a special test to get an idea if this simple requirement is fulfilled. As to larger samples, a simple chi-square goodness-of-fit test on a one-way category table can usually be used to reject the hypothesis about the uniform distribution of the last digits. If the test rejects, one should consider rounding the measurements to the next valid digit.

Decades of, say, daily sulfate concentration data may exhibit deviation from the uniform last digit distribution, however, because the laboratory and measuring equipment may have changed several times over time, for

example. As the accuracy of the equipment increased, so did the number of valid digits. If the new data records, with more valid decimals than the old ones, have just been added to the data stack, then upon a thorough analysis we may observe an increased frequency of zeros at the end of each number. A typical last digit frequency plot looks like the one below. Data for the above figure come from 135 years of monthly precipitation data from New York Park, c. f. U.S. Historical Climatology Network. The chart indicates, that a suspiciously large number of average daily temperatures ends with the digit 6.

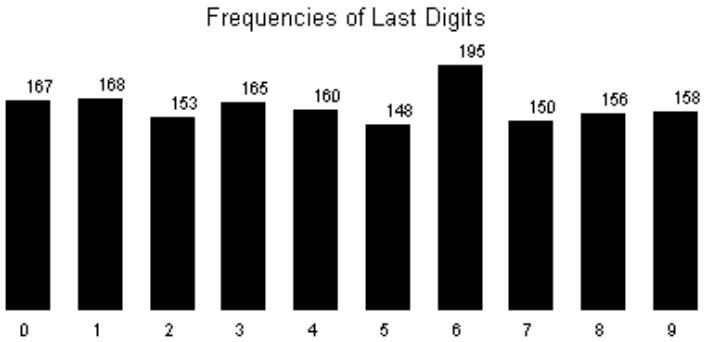

Figure *Last digit frequencies from 135 years of monthly precipitation data collected in New York Park.*

References

J. Mohapl: Measurement Diagnostics by Analysis of Last Digits. *Environmental Monitoring & Assessment*, pp. 407-441, 61, 2000.