I confirm that the following report and associated code is my own work, except where clearly indicated.


**Abstract**

The National Oceanic Atmospheric Association's (NOAA) mission is to understand changes to our climate, oceans, and weather.[1]Their mission statement also includes the importance of sharing the wealth of knowledge they find to the public by free access.[1] Glance (GCAG) is a component of NOAA that provides real-time monitoring data on global temperature anomalies from over the land and ocean.[2] Does an association between the temperature anomalies over the ocean versus over the land exist? Can the validity of a correlation test be trusted for performing such an analysis? The following report investigates under what scenarios a correlation test of mean temperature anomalies can be implemented using a simulation study. The dataset used in the study contains the global temperature anomalies and carbon emissions from 1751 to 2018. Global temperatures anomalies are recorded over land, ocean, and a mean combination of the two in Celsius (°C). The variable year was recorded in the current era (CE), and the annual carbon emissions were measured in millions of metric tons. The properties, such as mean and standard deviation for land and ocean is calculated to use as a reference value to resample (simulate) data. A parametric and non-parametric correlation test is conducted on the simulated data to retain the correlation coefficient and associated p-value. Simulation of the data is repeated a thousand times for scenarios of varying resampling properties. The size and power of each test are then retained as support for and potentially against the validity of conducting a parametric and non-parametric correlation test. The simulation study revealed that the validity of a correlation test can be trusted for the following temperature anomaly data, expect for under special circumstances.


**Introduction**

The National Oceanic Atmospheric Association's (NOAA) mission is to understand changes to our climate, oceans, and weather. Their mission also includes sharing this knowledge and information with the public.[1] An example of this information includes access to real-time monitoring data such as the Climate at a Glance (GCAG) .[2] GCAG is a component of NOAA that provides global temperature anomaly information, collected over the land and ocean, from the Global Historical Climatology Network-Monthly (GHCN-M) data set and International Comprehensive Ocean-Atmosphere Data Set (ICOADS).[2] Temperature anomalies are defined as the difference in measured temperature from a long-term average.[3] These anomalies can be used to provide us with an overview of how the global temperatures have changed over time.[3] Does an association between the temperature anomalies over the ocean versus over the land exist? Can the validity of a correlation test be trusted for performing such an analysis? The following report investigates under what scenarios a correlation test of mean temperature anomalies can be implemented using a simulation study in R version 4.1.1.[4]

**Methods**

The data used in this report is a blended product of two datasets from the GHCN-M and ICOADS. [2] A version of this dataset labeled as temp_carbon is available from the *dslabs* [5] package in R. It contains the global temperature anomalies and carbon emissions from 1751 to 2018. Global temperatures anomalies are recorded over land, ocean, and a mean combination of the two in Celsius (°C). The variable year was recorded in the current era (CE), and the annual carbon emissions were measured in millions of metric tons. Exploratory analysis is carried out to investigate the properties of the data, what distribution the data appears to follow, and whether the data set needs to be cleaned or rearranged for analysis. It is discovered that temperature observations are only recorded from 1880 to 2018. Since these are the parameters of interest a new data set containing no *NA* observations is required for simulation and analysis.

To analyze whether there is an association between land and ocean temperature anomalies recorded, a "Pearson" and "Spearman" correlation test is conducted. A Pearson test is the parametric case where the data is assumed to follow an independent normal distribution. The Spearman test is the non-parametric case where the data is free of a distribution. To investigate whether the validity of these tests can be trusted, new realistic data (i.e., simulated data) is randomly generated for the ocean/land temperature anomalies and tested for correlation at a 5% significance level, using the cor.test function with method type "Pearson" and "Spearman" in R. [4]

Simulation of the data is repeated a thousand times for scenarios of varying resampling properties. The size and power of each test are then retained as support for and potentially against the validity of conducting a "Pearson" and "Spearman" correlation test. Implementing these scenarios offer the capability of studying how well the test would perform if the true data varied from currently observed data. Each of the following scenarios are implemented: an increased/decreased sample size, altering the reference mean used in resampling, and changing the linear model coefficient of the dependent variable ( $\beta_1$ = land anomaly) to increase and decrease the distance from the intercept ( $\beta_0$).

In order to simulate data resembling the temp_carbon[5] dataset, the properties, i.e. mean and standard deviation, are needed as a reference value. Given that the association of variables is of interest, a linear model is fitted with a response variable of ocean temperature anomalies and a dependent variable of land temperature anomalies. From this linear model, the intercept and the dependent coefficient are retained to calculate the reference mean ($m\mu = \beta_0 + \beta_1 * land\ anomaly$) for simulating new data for ocean temperature anomalies. The reference mean, and standard deviation of the dependent variable are directly calculated using the mean and sd function in R. [4] These reference values are then set as the mean and standard deviation for simulating random deviates from a normal distribution. A normal distribution is chosen for resampling due to the apparent distribution of each variable present in the histograms produced during exploratory analysis.

Two functions, sim_fn_corr_pow and sim_fn_corr_size, were developed for this simulation study with arguments that can be specified to test for size and power under the different case scenarios. The function, sim_fn_corr_pow, was developed to determine the power of the test based on the alternative hypothesis that there is a significant correlation between ocean and land, being true. The function, sim_fn_corr_size, was developed to determine the size of the test based the null hypothesis that there is not significant correlation between ocean and land, being true.

## Results

For the scenario (1) in which the sample size of the simulated data is changed, the power of the Pearson and Spearman correlation test remains above 80% when there is sample size of at least eight observations for each group (Table 1). Figure 1 shows that a correlation, points in a diagonal pattern, is distinguishable with a sample size as small as eight and becomes clearly as sample size is increased. The size of the correlation test under this scenario remains below 6% even with a sample size three times larger the original sample size of n = 139 (Table 2).

| Sample.Size | Pearson.Coeff | Spearman.Coeff | Power.Pearson | Power.Spearman |
|---|---|---|---|---|
| 1000 | 0.918 | 0.911 | 1 | 1 |
| 300 | 0.919 | 0.91 | 1 | 1 |
| 8 | 0.905 | 0.856 | 0.97 | 0.865 |
| 6 | 0.9 | 0.84 | 0.872 | 0.579 |
| 5 | 0.899 | 0.835 | 0.765 | 0.283 |
| 4 | 0.876 | 0.813 | 0.485 | 0 |

*Table 1: Table of various sample size scenarios tested and the associated correlation coefficient and **power** for the parametric Pearson and non-parametric Spearman correlation test.*

| Sample.Size | Pearson.Coeff | Spearman.Coeff | Size.Pearson | Size.Spearman |
|---|---|---|---|---|
| 1000 | -0.00143 | 0.057 | 0.055 | NA |
| 300 | -0.00159 | 0.045 | 0.047 | NA |
| 8 | -0.0026 | 0.047 | 0.035 | NA |
| 6 | -0.0163 | 0.049 | 0.029 | NA |
| 5 | 0.0199 | 0.05 | 0.019 | NA |
| 4 | -0.003 | 0.059 | 0 | NA |

*Table 2: Table of various sample size scenarios tested and the associated correlation coefficient and **size** for the parametric Pearson and non-parametric Spearman correlation test.*

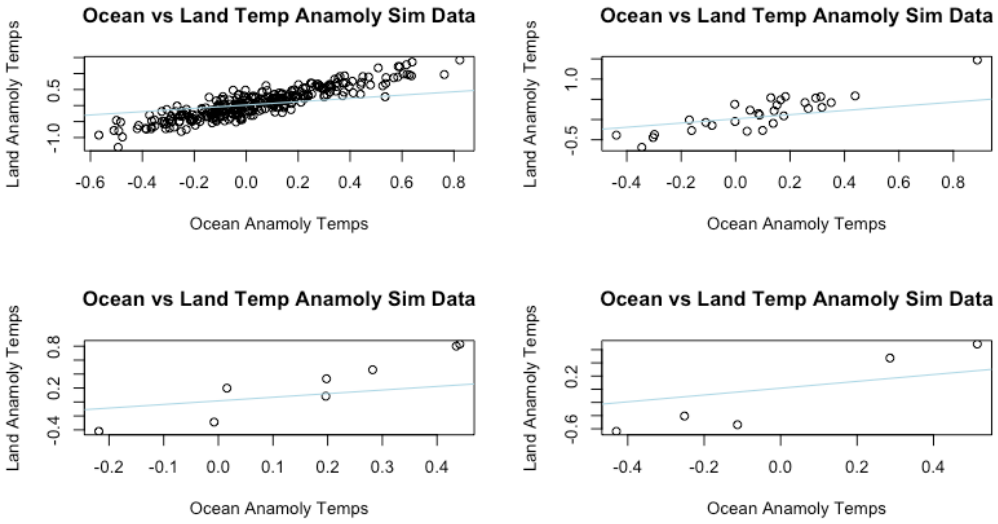


*Figure 1: Scatter plots of ocean temperatures vs land temperature anomalies for various sample sizes (n = 300, 30, 8, 5), with a reference line indicating the best fit of the temp_carbon [5] dataset.*

For the scenario (2) of altering the references mean parameter used in resampling of ocean/land temperature anomalies, the size of the Pearson correlation test appears to fluctuate about 5%. This occurs regardless of the mean parameter being larger or smaller than the reference mean (Table 4). While the power of both the Pearson and Spearman correlation test falls below a substantial power of 80% as the reference mean parameter continues to shrink to a value smaller than an eleventh the size of the reference mean. The associated power of shrinking the reference mean can be viewed in Table 3.

| Mu.Mulitplied.By | Pearson.Coeff | Spearman.Coeff | Power.Pearson | Power.Spearman |
|---|---|---|---|---|
| 0.0667 | 0.15 | 0.143 | 0.43 | 0.386 |
| 0.0909 | 0.207 | 0.198 | 0.708 | 0.673 |
| 0.1 | 0.23 | 0.22 | 0.777 | 0.736 |
| 0.2 | 0.419 | 0.403 | 1 | 0.999 |
| 0.5 | 0.759 | 0.74 | 1 | 1 |
| 5 | 0.996 | 0.995 | 1 | 1 |

*Table 3: Table of altered reference mean and the associated correlation coefficient and **power** for the parametric Pearson and non-parametric Spearman correlation test.*

| Mu.Mulitplied.By | Pearson.Coeff | Spearman.Coeff | Size.Pearson | Size.Spearman |
|---|---|---|---|---|
| 0.0667 | -0.00281 | 0.057 | 0.058 | NA |
| 0.0909 | 0.000532 | 0.055 | 0.047 | NA |
| 0.1 | 0.00343 | 0.042 | 0.042 | NA |
| 0.2 | 0.000115 | 0.062 | 0.057 | NA |
| 0.5 | 0.00193 | 0.045 | 0.053 | NA |
| 5 | -0.000491 | 0.045 | 0.054 | NA |

*Table 4: Table of altered reference mean and the associated correlation coefficient and **size** for the parametric Pearson and non-parametric Spearman correlation test.*

For the scenario (3) in which the coefficient of the depend variable ( $\beta_1 =$ land anomaly) in the linear model is altered, a minimum distance around 0.04 from the intercept of $\beta_0$ is required to maintain a substantial power of 80% for the Pearson and Spearman correlation test. If the distance between $\beta_1$ and $\beta_0$ continues to shrink, the power of both tests drastically drops when the distance between coefficients is less than or equal to 0.03. The associated power drop given the distance between coefficients can be viewed in Figure 1 and 2.

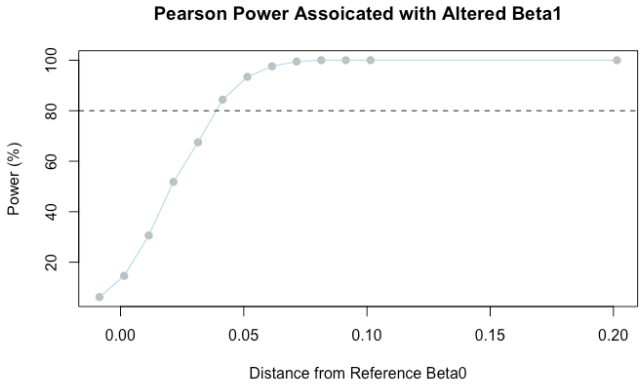


*Figure 3: Plot of the associated power for a Pearson correlation test as the distance from $\beta_0$ shrinks due to an altered $\beta_1$.The dashed line represents the 80% power threshold.*

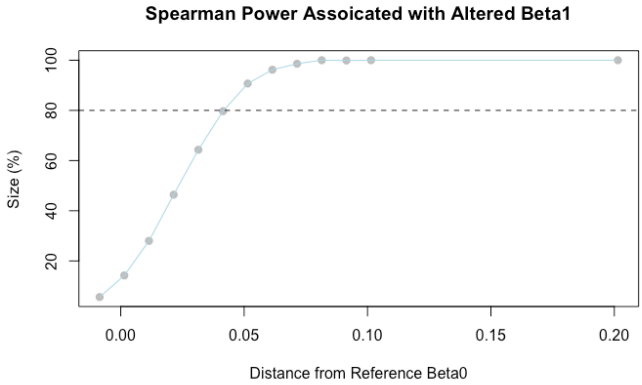


*Figure 4: Plot of the associated power for a Spearman correlation test as the distance from $\beta_0$ shrinks due to an altered $\beta_1$. The dashed line represents the 80% power threshold.*

## Conclusions

Scenario one implies that whether or not correlation truly exist for the temp_carbon[5] data set, as long as a minimum of eight observations are made for ocean and land temperature anomalies, the power and size would continue to fall within their respective significant ranges of 80 and 5%. Scenario two implies that if correlation does not exist, the size of the correlation test is essentially unaffected by a mean value that is larger or smaller than the reference mean. However, the power of the test starts to fall below the substantial threshold of 80% as the reference mean is an eleventh the mean value. Scenario three implies a correlation test is valid if the difference in linear model intercept and slope of the dependent coefficient is greater than 0.03.

The validity of a correlation test can be trusted overall for the following temperature anomaly data. There are however certain circumstances in which a correlation test would not be wise to use. If less than eight of the observations from the temperature anomalies were found to be valid. Meaning that the rest of the observations were incorrect due to sampling error and therefore would be discarded. If the true mean for land and ocean temperature anomalies is eleventh the size of the current means. Lastly if the intercept $\beta_0$ and dependent variable coefficient $\beta_1$ were found to differ by 0.03 or less.

## Bibliography

1. "About Our Agency." *About Our Agency | National Oceanic and Atmospheric Administration*, https://www.noaa.gov/about-our-agency.
2. NCEI.Monitoring.info@noaa.gov. "Climate at a Glance." *National Climatic Data Center*, https://www.ncdc.noaa.gov/cag/global/background.
3. Sanchez-Lugo. "Global Surface Temperature Anomalies." *National Climatic Data Center*, https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php.
4. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
5. Rafael A. Irizarry and Amy Gill (2021). dslabs: Data Science Labs. R package version 0.7.4. https://CRAN.R-project.org/package=dslabs

**Appendix 1**

Simulating data based on a Linear model from the original data set: