ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/306376553

Spurious Correlations in Time Series Data: A Note

Article *in* SSRN Electronic Journal · August 2016

DOI: 10.2139/ssrn.2827927

CITATIONS
17

READS
673

1 author:

Jamal Munshi
Sonoma State University
118 PUBLICATIONS **403** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

carbon cap and trade markets View project

All content following this page was uploaded by Jamal Munshi on 10 July 2019.

The user has requested enhancement of the downloaded file.

SPURIOUS CORRELATIONS IN TIME SERIES DATA: A NOTE

JAMAL MUNSHI

ABSTRACT: Unrelated time series data can show spurious correlations by virtue of a shared drift in the long term trend. The spuriousness of such correlations is demonstrated with examples. The SP500 stock market index, GDP at current prices for the USA, and the number of homicides in England and Wales in the sample period 1968 to 2002 are used for this demonstration. Detrended analysis shows the expected result that at an annual time scale the GDP and SP500 series are related and that neither of these time series is related to the homicide series. Correlations between the source data and those between cumulative values show spurious correlations of the two financial time series with the homicide series. These results have implications for empirical evidence that attributes changes in temperature and carbon dioxide levels in the surface-atmosphere system to fossil fuel emissions[1].

1. INTRODUCTION

Correlation can serve as evidence of causation only in controlled laboratory experiments. A correlation between field data taken under ambient and uncontrolled conditions does not, by itself, support causation. For example, a correlation between field data time series λ and μ could mean that λ causes μ, or that μ causes λ, or that a third unobserved variable causes both μ and λ, or even that the observed correlation does not contain information about causation (Wright, 1921) (Duesberg, 1989) (McArthur, 1980). Yet, such a correlation is nevertheless a precondition to a causation theory because in the absence of a correlation the data are inconsistent with direct causation (Wright, 1921).

This principle plays an important role in the study of anthropogenic global warming and climate change (AGW) where the causal agent is taken to be the emission of carbon dioxide from the burning of fossil fuels and its proposed effects are taken to be changes in atmospheric and oceanic carbon dioxide on an annual time scale (IPCC, 2007) (IPCC, 2014) (NOAA, 2015) and changes in surface temperature on a decadal time scale (Ricke-Caldeira, 2014). In this kind of empirical evidence, the theoretical time scale at which the proposed causation works must match the time scale of the correlation (Box, 1994). Thus a correlation between emissions and changes in atmospheric and oceanic carbon dioxide must exist at an annual time scale and a correlation between emissions and warming must exist at a decadal time scale.

An additional consideration in the use of correlation as empirical evidence is that long term trends in the two series that are unrelated to the causation theory being tested can create spurious correlations between them (Box, 1994). This effect is magnified if cumulative values are used instead of the source data (Munshi, 2016). To remove this effect, the two series are detrended and a correlation between the detrended series at the appropriate time scale is considered (Chatfield, 1989) (Prodobnik, 2008). In this short note we demonstrate these correlation issues with three time series and three different correlation pairs only one of which contains a rationale for correlation.

---

[1] August 2016, Updated 10/12/2016
Key words and phrases: climate change, global warming, carbon dioxide, fossil fuel emissions, carbon emissions, applied statistics, spurious correlations, cumulative values, spurious correlations between cumulative values, causation and correlation
Author affiliation: Professor Emeritus, Sonoma State University, Rohnert Park, CA, 94928 munshi@sonoma.edu

2. DATA AND METHODS

Values of the SP500 index on January 1 of each year from 1968 to 2002 are provided by the multipl.com website that tracks P/E ratios (MULTIPL, 2016). Annual GDP data for the USA at current prices are provided by the World Bank's Databank service (WorldBank, 2016). Data for the number of homicides per year in England and Wales are provided by the Official Statistics service of the government of the UK (GOV.UK, 2016). The sample period for all three data series was somewhat arbitrarily[2] set to 1968-2002.

The three time series in the study are labeled as SP5 (SP500), GDP (GDP), and HOM (Homicides). Three paired correlations are possible. They are GDP-SP5, GDP-HOM, and SP5-HOM. The correlation for each of these pairs is computed using three different methods. They are: (1) SOURCE = correlation between the annual source data as received, (2) CUMULATIVE = correlation between the cumulative values, and (3) DETRENDED = correlation between the detrended values.

Correlations are computed with the CORREL() function of Microsoft Excel. The OLS linear regression parameters for each series against time in years are computed with the LINEST() function of Excel and the detrended series is computed as the residuals of the LINEST() linear model over the entire sample period. The standard deviation of correlation is estimated using Bowley's method (Bowley, 1928) and the observed sample correlations are tested for statistical significance using the hypothesis test shown below.

$$H_0: \rho=0 \text{ against } H_A: \rho\neq 0$$

The hypothesis tests are carried out at a maximum false positive error rate of $\alpha=0.001$ in keeping with "Revised standards for statistical evidence" published by the National Academy of Sciences (Johnson, 2013). Since three correlations are computed for three variable pairs, nine different comparisons made. The overall study-wide false positive error rate is estimated as 0.009 in accordance with Holm's multiple comparison procedure (Holm, 1979). This means that there sa 0.9% chance of at least one false positive in nine comparisons.

The results for each of the three variable pairs are presented in Section 3 using four charts and one table. The first chart presents the source data. To make very different absolute values fit on the same chart, the absolute values are normalized as multiples of their value in 1968. The other three charts display a graphical depiction of the correlation for each of the three computational methods used. The table presents the numerical value of these correlations as well as t-tests for their statistical significance.

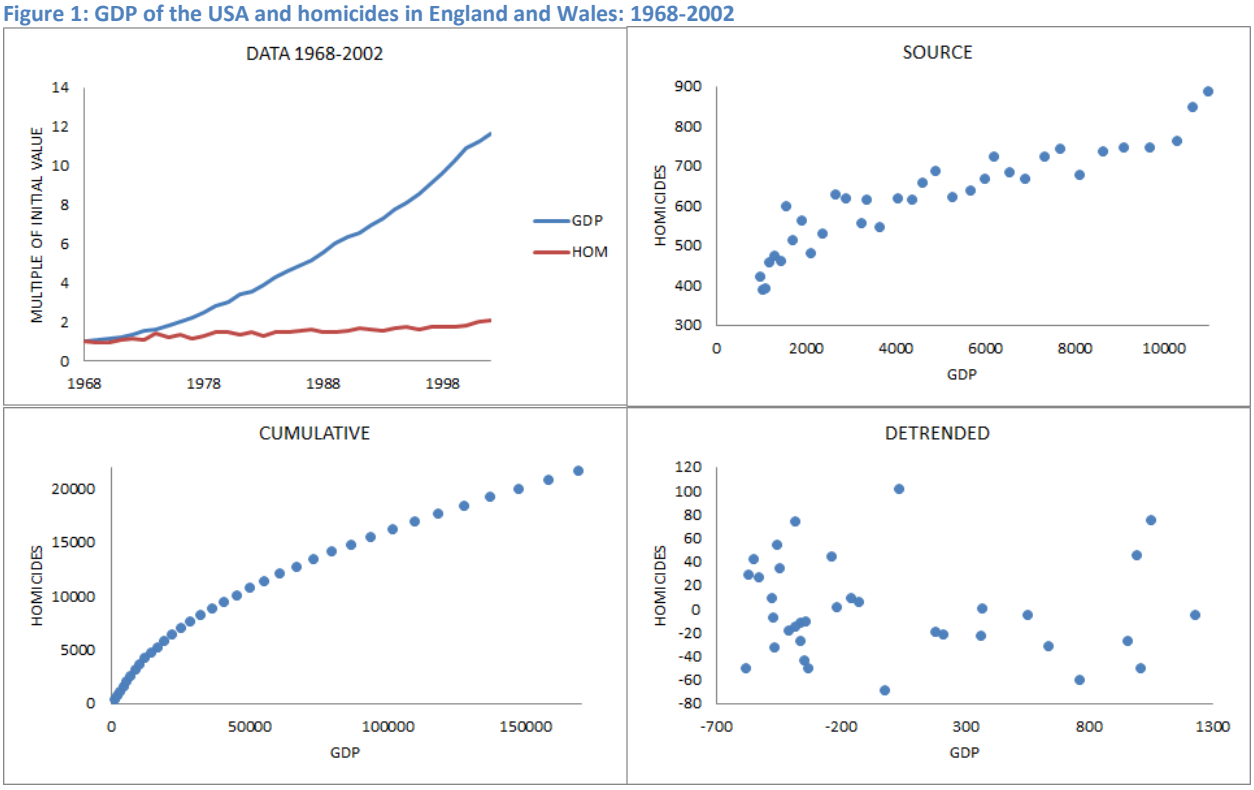
The use of correlations in the empirical data for the attribution of changes in surface-atmosphere properties to fossil fuel emissions is discussed in Section 4 in light of the results presented in Section 3. All data and computational details may be downloaded from the online data archive for this paper made available on Google Drive (Munshi, Correlation paper archive, 2016).

[2] This period contains a rising trend in the homicide series.

3. DATA ANALYSIS

3.1 GDP-Homicide

Figure 1: GDP of the USA and homicides in England and Wales: 1968-2002



Annual GDP of the United States and the number of homicides per year in England and Wales are displayed in the first frame of Figure 1 as multiples of their values in 1968. We note here that both series show an upward drift in time and demonstrate that the shared direction of the drift in time between the two time series creates spurious correlations between the source data and their cumulative values. Visual depictions of the correlation between the two annual time series appear in the next three frames. The SOURCE frame depicts the correlation between the source data, the CUMULATIVE frame depicts the correlation between their cumulative values, and the DETRENDED frame depicts the correlation between their detrended values.

Table 1 contains the numerical values of these correlations along with hypothesis tests for a positive correlation between the two variables in the form of t-tests for the null hypothesis H_0: $\rho>0$ where ρ is the correlation in the population from which the sample was taken. Both the SOURCE and CUMULATIVE correlations are high and statistically significant with the cumulative values showing the stronger correlation. By contrast, no evidence of a correlation between the DETRENDED series is found. These results imply that the SOURCE and CUMULATIVE correlations derive from a common but unrelated upward drift in values that does not necessarily imply a relationship at an annual time scale. No correlation remains when the shared upward drifts are removed.

These results are consistent with the absence of a conceptual basis for a relationship between the two series. We have no reason to believe that the homicide rate in England and Wales should be related to the GDP of the USA. As noted in a prior work, the correlation between cumulative values generates an anomalous and spurious correlation under certain conditions particularly when the two series contain an incidental drift in time in the same direction (Munshi, 2016). Such a correlation, though it may be very strong, does not contain information about the behavior of the two time series at an annual time scale. An additional consideration is that a reduction in degrees of freedom occurs when cumulative values are used because the same data appear numerous times in the computation of cumulative values. The figures in column 3 of Table 1 do not reflect this reduction in degrees of freedom.

Table 1: t-test for correlations between GDP and Homicide

GDP-HOM	SOURCE	DETRENDED	CUMULATIVE
R	0.9229	-0.0851	0.9770
R2	0.8518	0.0072	0.9545
N	35	35	35
SIGMA	0.0670	0.1734	0.0371
T-STAT	13.77	0.49	26.33
PVAL	0.00000	0.31337	0.00000

3.2 SP500-Homicide

Figure 2: SP500 stock market index and homicides in England and Wales: 1968-2002

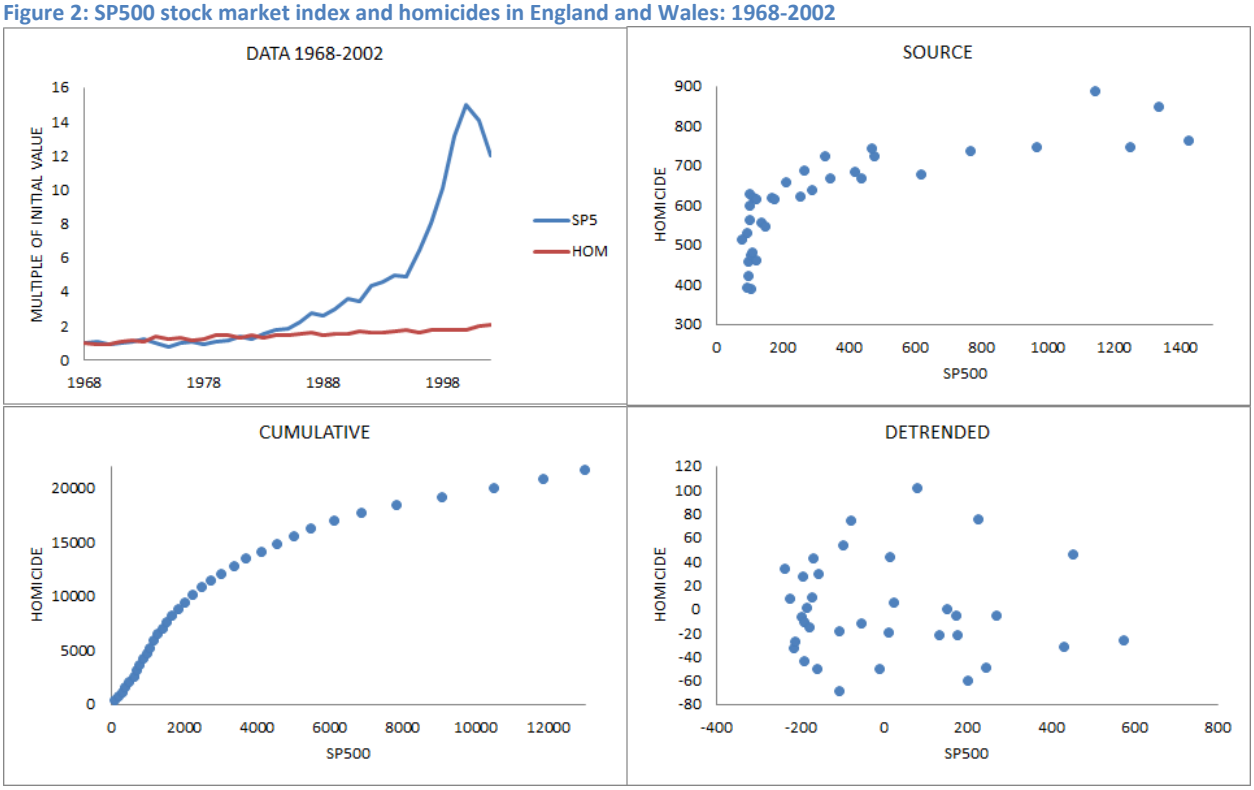


Table 2: t-test for correlations between SP500 and Homicide

SP5-HOM	SOURCE	DETRENDED	CUMULATIVE
R	0.7797	-0.0327	0.9224
R2	0.6079	0.0011	0.8509
N	35	35	35
SIGMA	0.1090	0.1740	0.0672
T-STAT	7.15	0.19	13.72
PVAL	0.00000	0.42601	0.00000

As in the case with GDP, we find that the SP500 stock market index is correlated with homicides in England and Wales 1968-2002 in the SOURCE data and in the CUMULATIVE values. However, no correlation can be detected in the DETRENDED series at an annual time scale. We consider the SOURCE and CUMULATIVE correlations to be anomalous and spurious and ascribe them to the effect of unrelated shared drifts in the long term trends in the two time series. We conclude that the data do not provide evidence that the SP500 time series is related to the number of homicides in England and Wales at an annual time scale net of long term trends. The results are consistent with the absence of a rationale for a relationship between the SP500 stock market index in the USA and homicides in England and Wales. The reduction in degrees of freedom that occurs when cumulative values are used is not reflected in the values shown in column 3 of Table 2.

3.3 GDP-SP500

Figure 3: GDP of the USA and the SP500 stock market index: 1968-2002

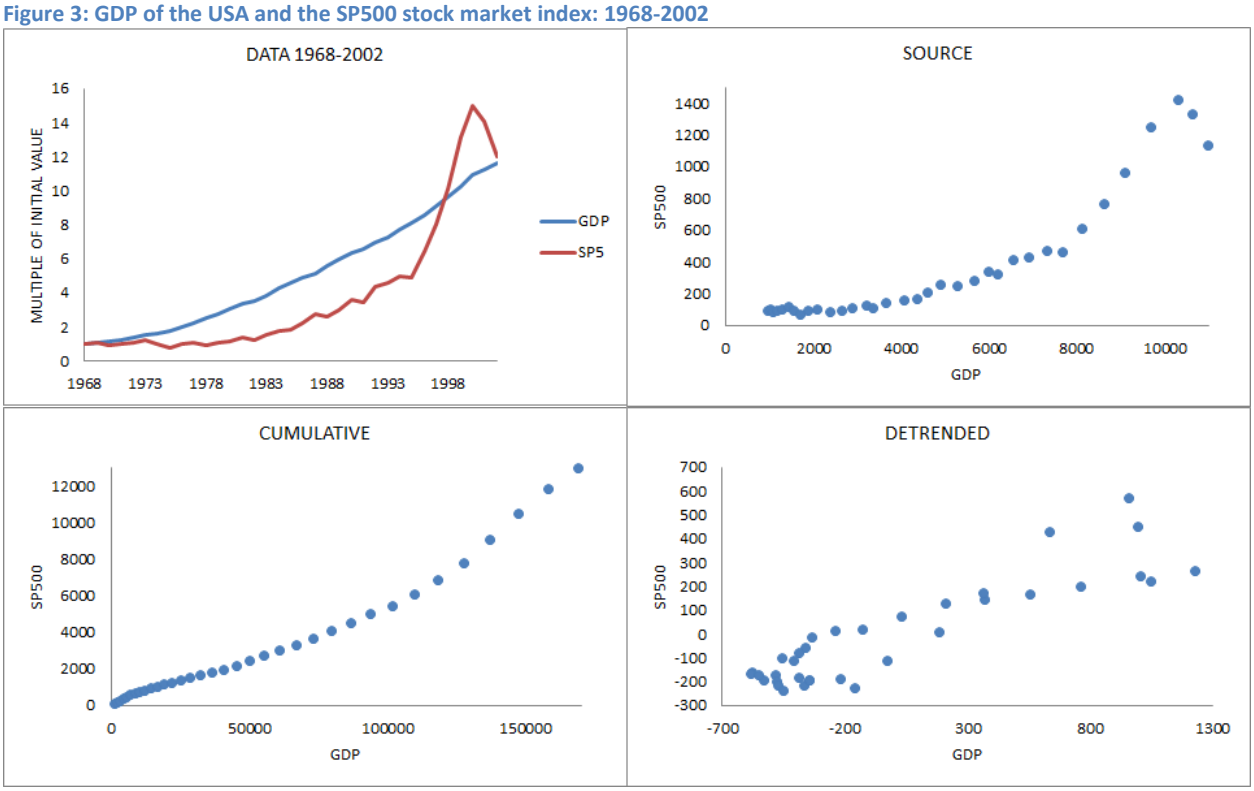


Table 3: t-test for correlations between GDP and SP500

GDP-SP5	SOURCE	DETRENDED	CUMULATIVE
R	0.9087	0.8991	0.9783
R2	0.8257	0.8083	0.9570
N	35	35	35
SIGMA	0.0727	0.0762	0.0361
T-STAT	12.50	11.80	27.10
PVAL	0.00000	0.00000	0.00000

Both of the financial variables GDP and SP500 reflect the underlying strength of the US economy and their close association is well known and well understood (EconompicData, 2015). This relationship is confirmed here in terms of a strong detrended correlation between GDP and the SP500 index at an annual time scale net of long term trends in the two time series. The reduction in degrees of freedom that occurs when cumulative values are used is not reflected in the values shown in column 3 of Table 3.

4. SUMMARY AND CONCLUSION

Unrelated time series data that share a common direction in their drift in terms of long term trend tend to show spurious correlations by virtue of their shared drift in time. In such cases, the effect of the common trend on correlation may be removed by de-trending the data and then computing the correlation between the detrended series (Chatfield, 1989) (Prodobnik, 2008) (Haan, 1977). The detrended correlation will show whether the movements of the two series net of trend are synchronized at the chosen time scale or whether they move independently.

In a demonstration of this procedure we show that the annual SP500 stock market index and annual GDP of the USA 1968-2002 show a strong correlation of R=0.9087 in the source data. Most of this covariance survives into the detrended series where we find that the detrended correlation is R=0.8991, a value that is statistically significant at a maximum false positive error rate of α=0.001 (Johnson, 2013). The result is consistent with the generally accepted relationship between these series (EconompicData, 2015). Both of these variables are reflective of the economic strength of the US economy and their strong correlation can be rationalized and understood in these terms.

The demonstration also shows the relationship between each of these financial series with the annual homicide rate in England and Wales 1968-2002. The source data show strong and statistically significant correlations between the financial time series and the homicide series with R=0.7797 for SP500 and R=0.9229 for GDP. However, the high levels of covariance observed in the source data do not survive into the detrended series where we find that the correlations are close to and statistically indistinguishable from zero. These results show that the high correlation in the source data are spurious and derive almost entirely from the common direction their drift in time. When that effect is removed the correlation spurious disappears.

Detrended analysis is unable to detect any relationship between the financial time series (SP500 and GDP) and the homicide rate. In fact we have no reason to believe that these variable pairs should be related in any way. Had there been a theory proposed for a causal relationship between the financial strength of the US economy and the homicide rate in England and Wales citing the strong correlations in the source data as empirical evidence, our results could be used to refute that claim by exposing the spuriousness of those correlations.

Spurious correlations of this nature are sometimes found in published research. For example, climate science attributes the rise in atmospheric carbon dioxide to fossil fuel emissions and cites correlations between the data as empirical evidence (IPCC, 2007) (IPCC, 2014) (Canadell, 2007) (Kheshgi, 2005). Detrended analysis shows that correlations between emissions and atmospheric CO2 and oceanic CO2 are spurious because they disappear when the data are detrended (Munshi, Responsiveness of Atmospheric CO2 to Anthropogenic Emissions, 2015) (Munshi, Fossil Fuel Emissions and Ocean Acidification, 2015). These anomalous results likely derive from large and perhaps unquantifiable uncertainties in natural flows of carbon dioxide in the surface-atmosphere system (Munshi, Uncertain Flow Accounting and the IPCC Carbon Budget, 2015).

An additional issue explored in this demonstration is the strong correlation between cumulative values shown in Tables 1, 2, and 3 as R = 0.9770, 0.9224, and 0.9783. In each case, the correlation between cumulative values is the highest observed correlation. Figures 1, 2 and 3 depict the exceptional smoothness of the relationship between cumulative values. In Tables 1 and 2 the correlations between cumulative values are shown to be spurious by the statistical insignificance of the detrended correlation. It is also noted that the use of cumulative values involves the repeated use of the same data and that greatly reduces the degrees of freedom in the computation of the correlation between cumulative values. The reduction in degrees of freedom to values close to unity (only the last value in the series is used only once) would make it difficult to establish statistical significance even for very high values of the correlation coefficient.

It has been demonstrated with Monte Carlo simulation that even for small tendencies for the two series to drift in the same direction, the cumulative values of random numbers tend to be correlated (Munshi, The spuriousness of correlations between cumulative values, 2016). This finding implies that the correlation between cumulative emissions and cumulative warming presented by climate science as empirical evidence for the attribution of warming to fossil fuel emissions (IPCC, 2007) (IPCC, 2014) (Allen, 2009) (Hansen, 1981) (Matthews, 2009) (Gillett, 2013) is spurious.

The demonstration of spurious correlations in this work shows that such correlations do not serve as empirical evidence in support of a causation theory because they are unreliable. All data and computational details used in this demonstration are available for download from an online data archive (Munshi, Correlation paper archive, 2016).

5. REFERENCES

Allen, M. (2009). Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, 458.7242 (2009): 1163-1166.

Anderson, K. (2011). Beyond 'dangerous' climate change: emission scenarios for a new world. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369.1934 (2011): 20-44.

Arora, V. (2011). Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters*, 38.5.

Bailey, I. (2011). Ecological modernisation and the governance of carbon: a critical analysis. *Antipode*, 43.3 (2011): 682-703.

Ballantyne, A. (2012). Increase in observed net carbon dioxide uptake by land and oceans during the past 50 years. *Nature*, 488.7409 (2012): 70-72.

Barnett, T. (1999). Detection and attribution of recent climate change. *Bulletin of the American Meteorological Society*, 80: 2631-2659.

Bodansky, D. (2001). The history of the global climate change regime. *International relations and global climate change*, (2001): 23-40.

Botzen, W. (2008). Cumulative CO2 emissions: shifting international responsibilities for climate debt. *Climate Policy*, 8.6, 569-576.

Bowley, A. (1928). The standard deviation of the correlation coefficient. *Journal of the American Statistical Association*, 31-34.

Box, G. (1994). *Time series analysis: forecasting and control.* Englewood Cliffs, NJ: Prentice Hall.

Callendar, G. (1938). The Artificial Production of Carbon Dioxide and Its Influence on Climate. *Quarterly Journal of the Royal Meteorological Society*, 64: 223-40.

Canadell, J. (2007). Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proceedings of the national academy of sciences*, 18866-18870.

Chatfield, C. (1989). *The Analysis of Time Series: An Introduction.* NY: Chapman and Hall/CRC.

Colglazier, W. (1991). Scientific uncertainties, public policy, and global warming: How sure is sure enough? *Policy studies journal*, 19.2 (1991): 61.

Draper&Smith. (1998). *Applied Regression Analysis.* Wiley.

Duesberg, P. (1989). Human immunodeficiency virus and acquired immunodeficiency syndrome: correlation but not causation. *Proceedings of the National Academy of Sciences*, 86.3 (1989): 755-764.

EconompicData. (2015). *Is there a relationship between the economy and the stock market?* Retrieved 2016, from EconompicData: http://econompicdata.blogspot.com/2015/04/is-there-relationship-between-economy.html

Falkowski, P. (2000). The global carbon cycle: a test of our knowledge of earth as a system. *Science*, 290.5490 (2000): 291-296.

Gillett, N. (2013). Constraining the ratio of global warming to cumulative CO2 emissions using CMIP5 simulations. *Journal of Climate*, 26.18 (2013): 6844-6858.

GOV.UK. (2016). *Historical crime data.* Retrieved 2016, from Official Statistics: https://www.gov.uk/government/statistics/historical-crime-data

Haan, C. T. (1977). *Statistical Methods in Hydrology.* Ames: Iowa State University Press.

Hansen, J. (1981). Impact of Increasing Atmospheric Carbon Dioxide. *Science*, 213: 957-66.

Hansen, J. (2016). Ice melt, sea level rise and superstorms:. *Atmos. Chem. Phys.*, 16, 3761–3812, 2016.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:2:65-70.

Hurst, H. (1951). Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 6: 770-799.

IPCC. (2007). *AR4 WG1 Chapter 7: Couplings between changes in the climate system and biogeochemistry.* Geneva: IPCC.

IPCC. (2014). *Climate Change 2013 The Physical Science Basis.* Geneva: IPCC/UNEP.

Johnson, V. (2013). *Revised standards for statistical evidence.* Retrieved 2015, from Proceedings of the National Academy of Sciences: http://www.pnas.org/content/110/48/19313.full

Kheshgi, H. (2005). Emissions and atmospheric CO2 stabilization: Long-term limits and paths. In *Mitigation and Adaptation Strategies for Global Change* (pp. 10.2 213-220).

Koutsoyiannis. (2003). Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences*, 48/1: 3-24.

Lacis, A. (2010). Principal Control Knob Governing Earth's Temperature. *Science*, 330.

Levin, I. (2000). Radiocarbon - a unique tracer of global carbon cycle dynamics. *Radiocarbon*, v42, #1, pp69-80.

Mandelbrot-Wallis. (1969). Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistival dependence. *Water Resources Research*, 5: 967-988.

Matthews, H. (2009). The proportionality of global warming to cumulative carbon emissions. *Nature*, 459.7248 (2009): 829-832.

McArthur, L. (1980). Illusory Causation and Illusory Correlation. *Personality and Social Psychology Bulletin* , 6.4 (1980): 507-519.

Meinshausen, M. (2009). Greenhouse-gas emission targets for limiting global warming to 2 C. *Nature*, 458.7242 (2009): 1158-1162.

MULTIPL. (2016). *S&P 500 Historical Prices by Year.* Retrieved 2016, from MULTIPL: http://www.multpl.com/s-p-500-historical-prices/table/by-year

Munshi, J. (2015). *Decadal Fossil Fuel Emissions and Decadal Warming.* Retrieved 2016, from ssrn.com/author=2220942: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2662870

Munshi, J. (2015). *Fossil Fuel Emissions and Ocean Acidification.* Retrieved 2016, from ssrn.com/author=2220942: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2669930

Munshi, J. (2015). *Responsiveness of Atmospheric CO2 to Anthropogenic Emissions.* Retrieved 2016, from ssrn.com: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642639

Munshi, J. (2015). *Uncertain Flow Accounting and the IPCC Carbon Budget.* Retrieved 2016, from ssrn.com/author=2220942: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2654191

Munshi, J. (2016). *Changes in the 13C/12C Ratio of Atmospheric CO2.* Retrieved 2016, from ssrn.com/author=2220942: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2781465

Munshi, J. (2016). *Correlation paper archive.* Retrieved 2016, from Google Drive: https://drive.google.com/open?id=0BxTCVvGifmvLSmtxaTgyLWdJUXM

Munshi, J. (2016). *Dilution of Atmospheric Radiocarbon CO2 by Fossil Fuel Emissions.* Retrieved 2016, from ssrn.com: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2770539

Munshi, J. (2016). *Seasonality and Dependence in Daily Mean USCRN Temperature.* Retrieved 2016, from ssrn.com: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2763358

Munshi, J. (2016). *The spuriousness of correlations between cumulative values.* Retrieved 2016, from ssrn.com: http://dx.doi.org/10.2139/ssrn.2725743

NOAA. (2015). *Ocean Carbon Uptake.* Retrieved 2016, from PMEL.NOAA: http://www.pmel.noaa.gov/co2/story/Ocean+Carbon+Uptake

Prodobnik, B. (2008). Detrended cross correlation analysis. *Physical Review Letters*, 100: 084102.

Revelle, R. (1956). *Carbon dioxide exchange between atmosphere and ocean and the question of an increase in atmospheric CO2 during the past decades.* UC La Jolla, CA: Scripps Institution of Oceanography.

Ricke-Caldeira. (2014). Maximum warming occurs one decade after carbon dioxide emission. *Environmental Research Letters*, V9 #12.

Robertson, I. (1887). Signal strength and climate relationships in 13C/12C ratios of tree ring cellulose from oak in southwest Finland. *Geophysical Research Letters*, 24.12 (1997): 1487-1490.

Roe, G. (2007). Why is climate sensitivity so unpredictable? *Science* , 318.5850 (2007): 629-632.

Solomon, S. (2009). Irreversible climate change due to carbon dioxide emissions. *Proceedings of the national academy of sciences*, pnas-0812721106.

WorldBank. (2016). *GDP.* Retrieved 2016, from The World Bank Data: http://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20.7 (1921): 557-585.

View publication stats