STUDY ON CORRELATION BETWEEN WIND SPEED AND TEMPERATURE USING INFORMATION ENTROPY

Amanda Goodrick

State University of New York at Binghamton, Binghamton, New York, USA

ABSTRACT

This project investigated temperature and wind speed correlation using information entropy. The hourly temperature and wind speed were retrieved over two weeks for two different regions of the United States, Boston, Massachusetts and Lincoln, Nebraska. Differential entropy was used to calculate mutual information between the hourly change in each variable. Findings revealed that mutual information shared between the two variables was greater in Lincoln than in Boston, indicating change in temperature and wind speed had a stronger correlation in Lincoln than in Boston.

KEYWORDS

Information entropy, differential entropy, mutual information

1. INTRODUCTION

Typical statistical methods have been used to analyze storm patterns and air pressure to investigate the relationship between temperature and wind speed, though these methods often do not incorporate certain factor relations such as nonlinear correlation. As an exercise in applied differential entropy, mutual information was used to analyze this relationship. Boston, MA and Lincoln, NE were chosen based on geographical location. Boston is in the northeast United States, near the ocean, and Lincoln is landlocked near the middle of the United States, far from the ocean.

2. KEY CONCEPTS

Data description and key concepts follow.

2.1. Data Description

Historical hourly temperature (in degrees Fahrenheit) and hourly wind speed (in kilometers per hour) were obtained for Boston, MA and for Lincoln, NE for two weeks from November 17th to December 1st [1][2]. Original data contained 359 observations per variable. To create stationarity and remove potential dependence among each hourly observation, first differences were used. The data analyzed was change in hourly temperature and hourly wind speed with 358 values per variable. Temperature and wind speed are continuous variables, so the differences are also continuous.

2.2. Differential Entropy

Differential entropy, also called continuous entropy, can be negative. This method uses integrals rather than summation as an attempt by Shannon to extend Shannon entropy. Using summation to calculate entropy with binned continuous data is quite sensitive to bin size, so integration is used to omit variation due to bin size. Differential entropy of a variable has little meaning on its own, but entropies can be compared.

2.3. Mutual Information

Mutual information of two variables is calculated by summing individual entropies and subtracting joint entropy. Mutual information, also known as mutual dependence, is always non-negative, and it provides the amount of information shared between two variables. Mutual information is used when linear correlation is weak, non-existent, or when data patterns indicate non-linear correlation.

3. METHODS

Scatterplots and linear correlations were investigated. To calculate differential entropy for raw data, a smoothing method estimates the actual data by fitting a smooth curve to the histogram of the data. The first method estimated the data with a normal distribution with the same parameters as the data. The probability density function (pdf) for each normal curve was used to calculate individual entropies with X as the temperature change and Y as the wind speed change.

Differential entropy of the variable X was calculated by $H(X) = -\int \text{pdf}(x)\text{Log}_2\big(\text{pdf}(x)\big)\,dx$. The joint pdf for two normally distributed variables was used to calculate mutual entropy. Mutual information, I(X;Y) was calculated with the formula I(X;Y) = H(X) + H(Y) - H(XY) where H(X) and H(Y) are individual entropies and H(XY) is the mutual entropy.

A kernel smoothing distribution provided an improved data fit [3]. This method fits a smooth distribution when the data does not follow a typical well-known distribution.

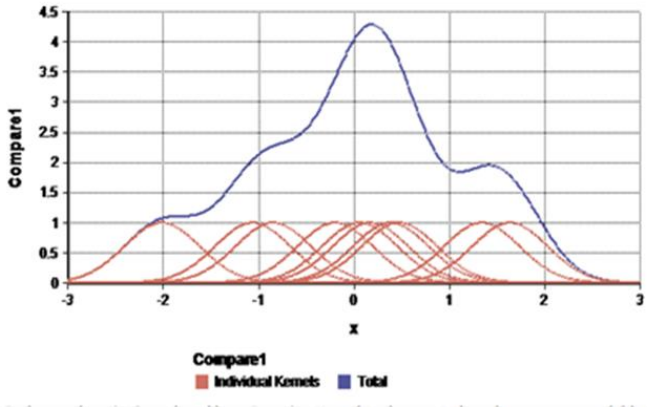

Figure 1. Kernel Smoothing.

The Kernel density estimation (KDE) replaces each point with a Gaussian, or normal, shaped kernel, and these kernels are added at each value to obtain the height of the smooth curve at each interval in the data range [4]. Figure 1 illustrates this process in an example from ten sample points and 0.4 bandwidth. Each data value is replaced with a standard normal distribution as

though the frequency of each value follows a normal distribution. For each interval of 0.4, the height of each normal curve is added at points in that interval. The height of the fitted curve in each interval fits the height of the histogram of actual values [3]. This process is typically done with symmetric distributions, and bandwidth choice could easily result in over fitting or under fitting.

Mathematica's Smooth Kernel Distribution provided the pdf for variable x, which was a linear interpolation of $\frac{1}{nh}\sum k\,(x - x_i/h)$ with sample size n, kernel function k, and bandwidth h [5]. Bandwidth was estimated by Silverman's rule, $\hat{h} = (4\hat{\sigma}^5/3n)^{1/5} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$ where $\hat{\sigma}$ is the standard deviation from the data [6]. The individual entropies were calculated from each KDE. The joint pdf, provided by Mathematica's Product Distribution, was used to calculate mutual entropy.

4. RESULTS

Descriptive statistics and linear correlation coefficients are in Table 1.

Table 1. Mean (Std. Dev.) and Correlation.

City	Temp Change (degrees)	Wind Speed Change (km/h)	Linear Correlation
Boston, MA	0.018 (1.6)	0.013 (1.8)	0.17
Lincoln, NE	-0.013 (2.8)	0.001 (2.2)	0.38

The variables have a higher correlation in Lincoln, though correlations are weak in both cities, verifying the need to use a different method to investigate the relationship. Patterns in Figure 2 indicate nonlinear correlation between the two variables in Boston.

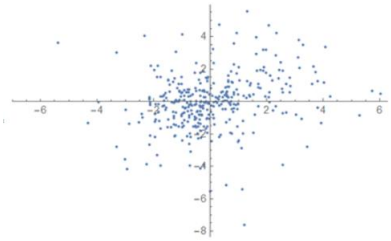

Figure 2. Scatterplot of Boston Data

Figure 3 illustrates normal distribution curves fit to the variables. This curve was visibly not the best estimate of the data.

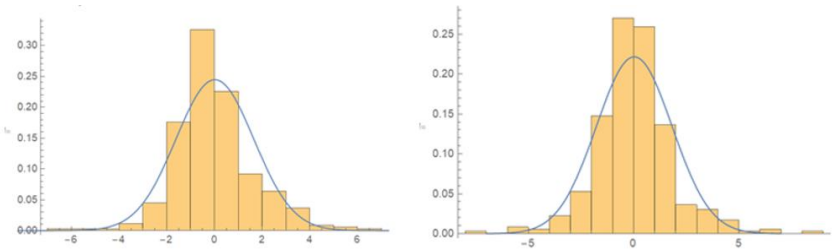

Figure 3. Normal Distribution Fit

In figure 4, the KDEs with bandwidths of 0.52 and 0.59 for temperature and wind speed, respectively, provided a closer fit to the data.

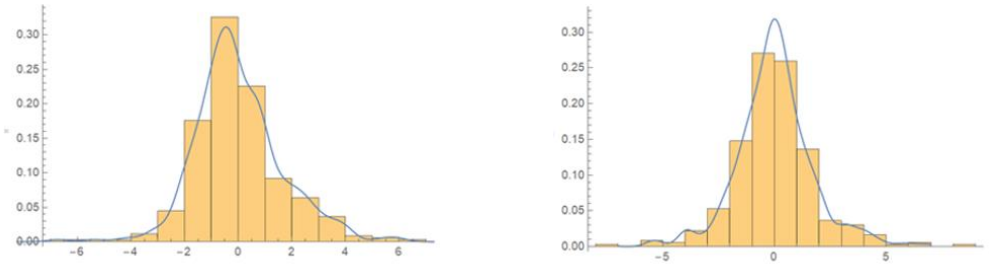

Figure 4. Kernel Smoothing Distribution Fit

Entropies were close to the values calculated with normal distributions. Mutual information indicated 0.4 bits of information shared between the two variables in Boston with the KDE.

The same analysis was conducted on temperature change and wind speed change in Lincoln, Nebraska. The linear correlation coefficient in Table 1 indicated a weak positive correlation, visible in figure 5 (a).

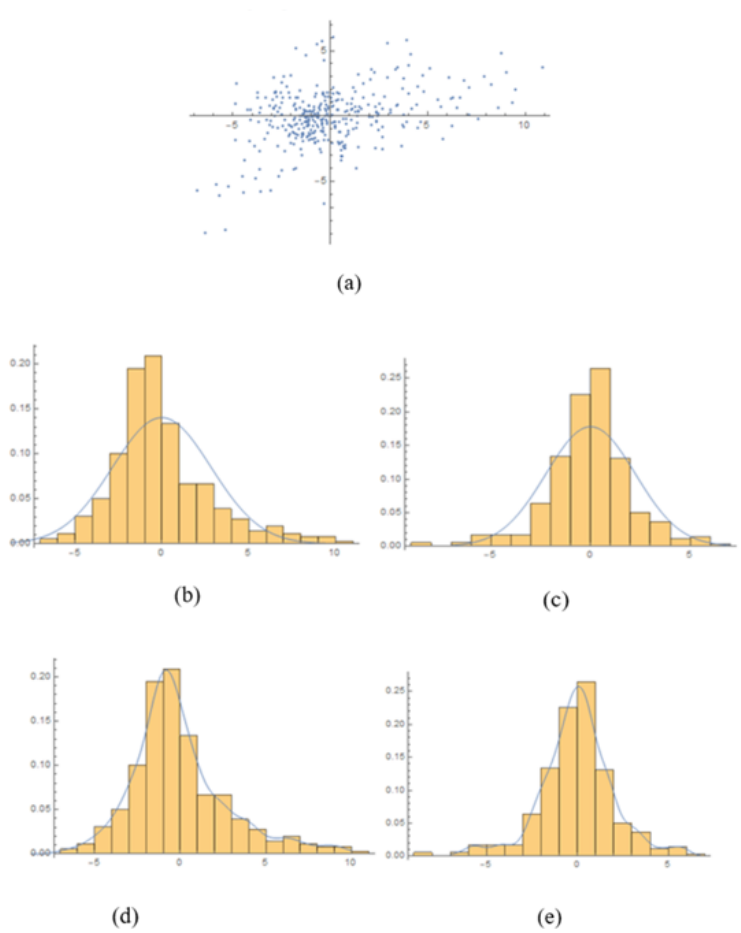

Figure 5. Lincoln, NE Temperature Change and Wind Speed Change

In Figure 5, plot (a) is the scatterplot of the temperature change vs wind speed change, and plots (b) and (c) are the histograms fitted with a normal distribution. Plots (d) and (e) are the histograms fitted with the KDE. KDE bandwidths were 0.92 and 0.72 for temperature and wind speed, respectively. Individual and joint entropies with mutual information from the normal distribution and from the KDE are in Table 2 and Table 3, respectively.

Table 2. Normal Distribution Fit Entropies.

City	Temp Change	Wind Speed Change	Joint Entropy	Mutual Info
Boston, MA	2.75321	2.89622	5.565	0.0844368
Lincoln, NE	3.55225	3.21102	6.24259	0.522885

Table 3. KDE Entropies.

City	Temp Change	Wind Speed Change	Joint Entropy	Mutual Info
Boston, MA	2.71385	2.81012	5.1476	0.376361
Lincoln, NE	3.45359	3.08149	5.21885	1.31623

Entropies indicate the same relationship between variables using either continuous distribution.

5. CONCLUSION

Results indicate more information was available in Boston with known temperature change than with known wind speed change. Information shared between the two variables was less than one bit. In Lincoln, more information was available with known wind speed change than with known temperature change. Mutual information was greater in Lincoln, suggesting that change in temperature and wind speed had a stronger relation in Lincoln than in Boston.

6. LIMITATIONS & FUTURE WORK

This analysis was limited by availability of data, which was only freely accessible for two week periods. Future work involves the same analysis with more variables, such as humidity and air pressure, additional geographical locations to determine impact of location on weather factors, and incorporating other smoothing estimators to determine if mutual information depends on the method.

REFERENCES

[1] Weather history download boston. (n.d.). Retrieved November 24, 2020, from https://www.meteoblue.com/en/weather/archive/export/boston_united-states-of-america_4930956

[2] Weather history download lincoln. (n.d.). Retrieved November 24, 2020, from https://www.meteoblue.com/en/weather/archive/export/lincoln_united-states-of-america_4930956

[3] Kernel density smoothing. (n.d.). Retrieved November 30, 2020, from https://wiki.analytica.com/Kernel_Density_Smoothing

[4] User484. (2015, May 01). How to evaluate differential entropy from raw data? Retrieved November 24, 2020, from https://mathematica.stackexchange.com/questions/65430/how-to-evaluate-differential-entropy-from-raw-data

[5] Smooth kernel distribution-Wolfram language documentation. (n.d.). Retrieved July 17, 2022, from https://reference.wolfram.com/language/ref/SmoothKernelDistribution.html

[6] *Physics 509: Kernel density estimation - phas.ubc.ca*. (n.d.). Retrieved July 17, 2022, from https://phas.ubc.ca/~oser/p509/Lec_23.pdf

AUTHORS

Amanda has a MS degree in Mathematics and a MS degree in Data Analytics. She is currently pursuing a PhD in Systems Science while teaching full-time at a 4-year university.