**Identifying Contextual Outliers in Weather Data**

Normal, calm weather varies from hour to hour, day to day, and month to month; however, extreme variations in weather can cause chaos and destruction. Developing methods to detect extreme variations in weather data could help scientists develop early warning systems that prepare and protect people from bad weather. This paper explores two methods for outlier detection: a nonparametric statistical approach that uses histograms, and a density-based proximity-based approach that uses a data object’s local outlier factor to determine whether or not it’s an outlier. The KY Mesonet dataset used in this paper encompasses six cities from a similar region in Kentucky. The cities are then ranked from least to most volatile during the months in which the data is recorded.

**Initial Observations about the Data**

The data is ordered and discrete. All of the data is captured every five minutes starting from March 1, 2016 to the end of May 31, 2016. There are six different sets of data from KY Mesonet. The six different sets of data are from different locations in Kentucky and are:

1. MRRY – Murray, KY
2. GRHM – Henderson, KY
3. PGHL – Hopkinsville, KY
4. HTFD – Hartford, KY
5. RSVL – Russellville, KY
6. FARM – Bowling Green, KY



Each location is described by 11 attributes. The attributes and their properties are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Abbreviation** | **Unit** | **Attribute Type** |
| Station ID | STID | Degrees | Nominal |
| Time | UTME | Time | Numeric - interval |
| Air temperature | TAIR | Celsius | Numeric - interval |
| Relative humidity | RELH | Percentage | Numeric - ratio |
| Dewpoint | TDPT | Celsius | Numeric - interval |
| Wind speed | WSPD | m/s | Numeric - ratio |
| Wind speed max gust | WSMX | m/s | Numeric - ratio |
| Wind direction | WDIR | degrees | Numeric - ratio |
| WD max | WDMX | degrees | Numeric - ratio |
| Precipitation | PRCP | mm | Numeric - ratio |
| Solar radiation | SRAD | W/m2 | Numeric - ratio |

Since each location has roughly the same amount of data, the base rate for each location is about 16.67%.

**Formulating the Outlier Detection Problem**

Outlier detection can be supervised, semi-supervised, or unsupervised. In supervised and semi-supervised methods, normal data are labeled by domain experts and the target variable indicates whether or not the data object is an outlier. However, due to the nature of outliers, there is a class imbalance which can cause degraded performance from the classification model. In unsupervised methods, the data objects do not have class labels, but are instead discovered through statistical approaches, proximity-based approaches, and clustering-based approaches.

This paper explores the unsupervised methods that relate to the statistical and proximity-based approaches. These methods are chosen because KY Mesonet data does not contain class labels for normal data. In addition, the weather data is from the same region in Kentucky and the time range spans a similar season of the year, so it is likely the normal objects are clustered and outliers in the whole dataset should be caused by similar forces. Further, contextual outliers will be found for each location.

The inputs and outputs for each unsupervised method are considered. The inputs for the statistical methods are the data points, the number of bins, and the size of each bin. The output is a histogram that acts as an outlier detection model for anything that falls into bins beyond a certain threshold. The inputs for the proximity-based approach are the data points and the value for k neighbors.

**Data Preprocessing**

KYMesonet data is preprocessed in three ways: first, the missing values of an attribute are imputed using the attribute’s mean; second, the time stamp is split into individual attributes which consist of month, day, year, hour, minute, and second; and third, the data is scaled by finding the mean and standard deviation of the entire column vector, then standardizing each data point. The range of the z-scores for each column is as follows:

|  |  |  |
| --- | --- | --- |
| Attributes | Min Z-Score | Max Z-Score |
| Month | -1.218226 | 1.218073 |
| Day | -1.675693 | 1.712497 |
| Hour | -1.661289 | 1.661364 |
| Minute | -1.593221 | 1.593252 |
| Second | -0.9999528 | 1.0000409 |
| TAIR | -2.881572 | 2.426686 |
| RELH | -2.483487 | 1.459472 |
| TDPT | -3.287754 | 2.383150 |
| WSPD | -1.356290 | 6.970738 |
| WDIR | -1.639387 | 1.775799 |
| WSMX | -1.455247 | 7.123193 |
| WDMX | -1.846467 | 1.767372 |
| SRAD | -0.6853444 | 4.0106989 |
| PRCP | -0.1079447 | 119.0781485 |

It appears that wind speed (WSPD), wind speed max (WSMX), solar radiation (SRAD), and precipitation (PRCP) have the most anomalous data points.

**Methods**

The 64bit version of R programming language is used to perform all of the computations. One library from R is used to discover the outliers. Training and test data sets are not required in this study since the study focuses on outliers rather than classification, though, using outliers to develop classification models would be a next reasonable step in the study.

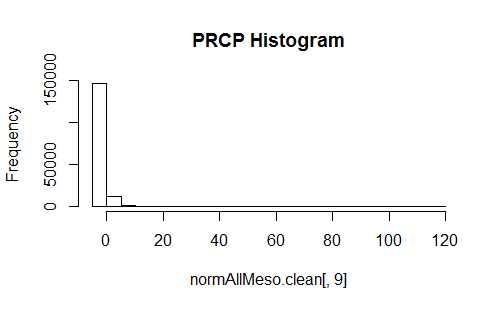
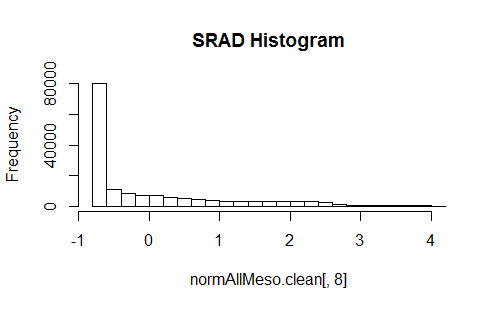
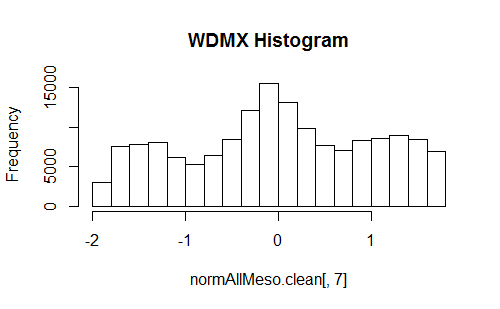
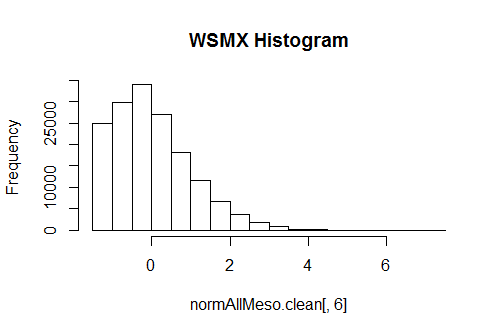
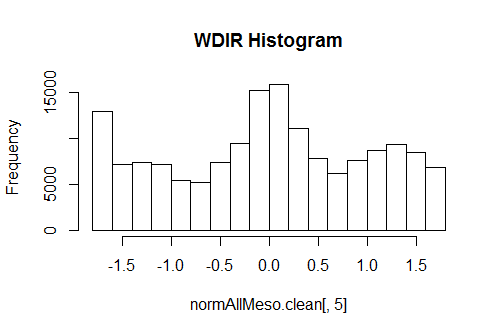
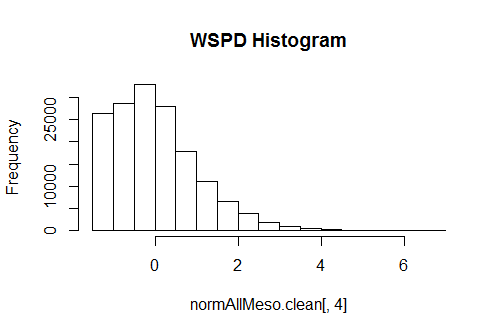
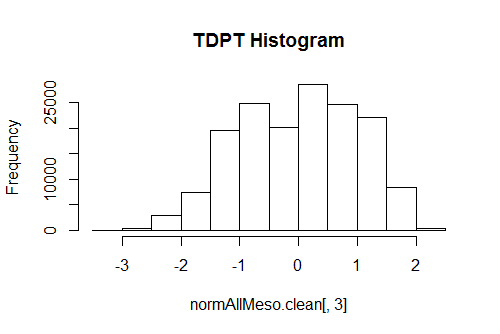
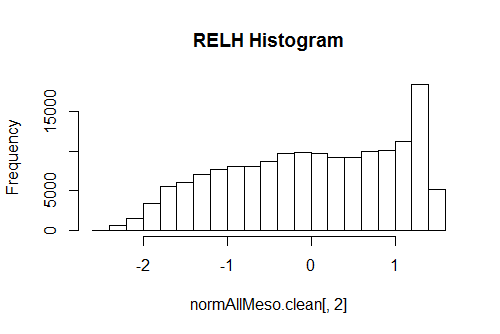
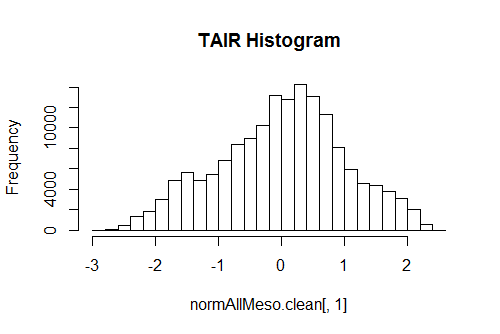
The data does not nee

**Statistical Outlier Detection Using Histograms**

The histogram is a nonparametric statistical outlier detection method that works by detecting values that fall outside of a designated range of binned values. Histograms are created for each attribute from the KY Mesonet dataset. The inputs for the histogram are a one dimensional vector and bin width. The output is the frequency of numbers found in the number range of each bin. The bin width was determined by looking at the five-number summary for each attribute. Since each attribute was normalized, the values from the five-number summaries represent the number of standard deviations a value is away from the mean. By the 68-95-99.7 rule, any outlier with a Z score greater than or equal to 3 is considered the top 0.3%, so data objects that fall in this range will be considered outliers.

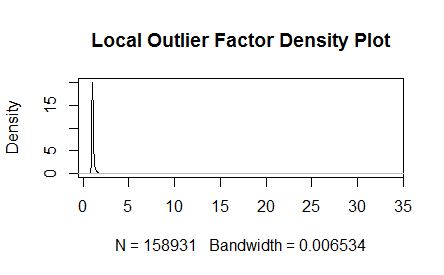
|  |  |
| --- | --- |
| **Attribute** | **Five-Number Summary for Normalized KY Mesonet Data** |
| TAIR | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -2.8820000 -0.6691000 1.000094 -0.0000033 0.6650000 2.4270000 |
| RELH | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -2.4830000 -0.8075000 1.000094 -0.0000156 0.9008000 1.4590000 |
| TDPT | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -3.288000 -0.809900 1.000093 -0.000017 0.788100 2.383000 |
| WSPD | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -1.356000 -0.748100 1.000078 0.000032 0.562300 6.971000 |
| WDIR | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -1.6390000 -0.8103000 1.000091 0.0000039 0.8347000 1.7760000 |
| WSMX | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -1.45500 -0.72950 1.000089 0.00002 0.58150 7.12300 |
| WDMX | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -1.8460000 -0.7292000 1.000094 0.0000002 0.8366000 1.7670000 |
| SRAD | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -0.685300 -0.685300 1.000094 -0.000001 0.395800 4.011000 |
| PRCP | Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.  -0.1079 -0.1079 1.000094 0.0000 -0.1079 119.1000 |

The attribute TAIR has a rather normal distribution. The bins chosen to represent the outliers are less than -**2.5** and greater than **2.5.** The attribute RELH has a distribution with a heavy left tail, as indicated by the minimum score being an entire standard deviation different than the absolute value of the maximum score. Therefore, the bins chosen to represent the outliers are less than -**2.25** and greater than **1.25**. The attribute TDPT also has a distribution with a heavy left tail, as indicated by the minimum score being an entire standard deviation different the absolute value of the maximum score. Therefore, the bins chosen to represent the outliers are less than **-3** and greater than **2**. The attribute WSPD has a distribution with a heavy right tail, as indicated by the maximum score being over five standard deviations greater than the absolute value of the minimum score. Therefore, the bins chosen to represent the outliers are less than the -**1** and greater than **3**. The attribute WDIR is rather normally distributed because the absolute values of both minimum and maximum values are similar. The bins chosen to represent the outliers are less than **-1.25** and greater than **1.25**. The attribute WSMX has a distribution with a heavy right tail, as indicated by the maximum score being over five standard deviations from the absolute value of the minimum. Therefore, the bins chosen to represent the outliers are less than **-1** and greater than **3**. The WDMX attribute is rather normally distributed because the absolute values of both the minimum and maximum values are similar. The bins chose to represent the outliers are less than **-1.25** and greater than **1.25**. The attribute SRAD has a distribution with a heavy right tail, as indicated by the maximum score being over three standard deviations from the absolute value of the minimum value. Therefore, the bins chosen to represent the outliers are less than -**0.25** and greater than **3**. Finally, the attribute PRCP has a distribution with a heavy right tail, as indicated by the maximum score being nearly 119 standard deviations away from the absolute value of the minimum score. Therefore, the bins chosen to represent the outliers are less than **0.5** and greater than **3**. The histograms below illustrate the bins for the data.



**Proximity-Based Outlier Detection Using Local Outlier Factor**

The local outlier factor (LOF) scores were computed in R using the Data Mining with R2 (DMwR2) library. The inputs for the LOF algorithm were the normalized KY Mesonet dataset and the value for k. For this algorithm, the test values for k were 3, 4, and 5. The computation took 9038 seconds to compute for k = 5, 8741 seconds for k = 4, and 8203 seconds for k = 3. The outliers were less frequent with k = 5, thus k = 5 is used for the dataset. The density plot for k = 5 is shown below.

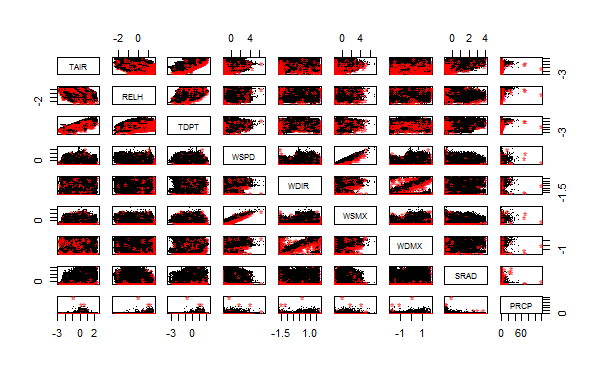


The five-number summary for the LOF scores are as follows:

Min. 1st Qu. Std. Dev. Mean 3rd Qu. Max.

0.8526 1.0010 0.20177 1.0820 1.1080 33.6900

Since the LOF scores were distributed with a heavy right tail, the outliers chosen contained LOF scores greater than the mean + (std. dev. multiplied by 3). A plot of the density-based outliers is shown below. The outliers are labeled in red.



It is evident from the plot above that density-based outliers occur within what seems to be the general group of values. This can be explained due to the fact that multiple locations are included in the density plot.