A COMPARATIVE ANALYSIS OF UNIFORMITY TESTS IN CIRCULAR STATISTICS

A Thesis

by

ARMANDO MARCIAL RUIZ MORA

BBA, Universidad Autonoma del Noreste, Mexico, 2009

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

Texas A&M University-Corpus Christi
Corpus Christi, Texas

December 2018

A COMPARATIVE ANALYSIS OF UNIFORMITY TESTS IN CIRCULAR STATISTICS

A Thesis

by

ARMANDO MARCIAL RUIZ MORA

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

JOSE GUARDIOLA, PhD
Chair

LEI JIN, PhD
Co-Chair

BLAIR STERBA-BOATWRIGHT, PhD
Committee Member

December 2018

ABSTRACT

In the context of circular statistics, data may not behave uniformly around the circle, exhibiting a preferred direction, thus the need to find appropriate methods to detect departures from a uniform distribution. First, we discuss some uniformity tests for circular data. Then, a likelihood ratio test (LRT) is proposed using standard statistical theory. The cardioid distribution seems particularly adequate as an alternative hypothesis for the LRT when data show a smooth transition in a unimodal preferred direction. Second, when the data are not uniformly distributed, we apply different circular regression methods to devise the patterns of dependence on some independent variables. When the distributional assumptions for parametric regression analysis are violated, a bootstrap method is proposed to test the regression coefficients. Finally, we have applied the uniformity tests and the circular regression methods to analyze wind directional data. Numerical results are summarized in tables for comparison purposes. We observe consistent results that the wind direction is not uniformly distributed via the uniformity tests. For the regression methods, we notice that wind speed is not significant to predict direction while the pattern of wind direction depends on the circular variable time of day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES PAGE

CHAPTER I: INTRODUCTION

Circular statistics (Jammalamadaka, 2001) is an area within statistics concerned with analyzing data being plotted on the unitary circle. In the aforementioned context, we consider statistical inference including the detection of deviations from uniformity, von Mises distributions in an appropriate manner and bootstrap tests for the circular regression.

1.1 Circular Statistics

A branch of statistics that studies the techniques necessary to analyze data that has circular or cyclic origin is called circular (or directional) statistics. Circular observations can be plotted on a circle with radius one. Circular statistics has a wide range of possibilities for analysis, connecting both circular and linear data. There are various scientific areas that can benefit from it.
In physics, obtaining rotation measurements from circular spectral polarization data, this is achieved using the maximum likelihood for a distribution with no fluctuation under coordinate transformation. The von Mises distribution works for this purpose. In Medicine, the analysis and classification of heart rhythm with electrocardiogram-waves is achieved with the readings of the periodicity of signals, extended through out segments with no intersection and R-waves observations on the unit circle. In chemistry, detecting uniformity of data in chemical processes such as in a research for decreasing effectiveness of electrolysis and flow reactor with different increased temperatures in a tank has been performed by applying three procedures for this purpose in circular statistics, such as Rao's spacing test, Kuiper's V-test and Rayleigh's test (Mardia, 2000). In biology, commonly animal behavior has a unimodal departure from a uniform distribution, but the multimodal case is also possible. For instance, it is known that domestic cats are active during the night, with a null hypothesis of uniformity for the activity of cats and an alternative hypothesis stating the opposite, because we know that cats have two peaks of activity, in the morning and in the evening our results should be able to show a multimodal distribution.

More recently in probabilistic machine learning, there has been going on some research trying to connect circular statistics to machine learning, topics that traditionally have not been researched together, for example applying models and approximate inference algorithms, using the Multivariate Generalized von Mises distribution that is the equivalent Gaussian distribution on the circle.

Circular statistics originated from the necessity to represent directions but also to obtain meaningful statistical analysis and inference. For the two-dimensional case, directions are plotted as points on the unitary circle, these are are called circular data. Three-dimensional unit vectors are represented by two angles, or observations on the unit sphere, they are also known as spherical data and even multivariate data can also be represented on the unit hypersphere. Our research will be based on the two dimensional case.

There are different manners to represent circular data. Let us focus on the most important ones, the compass and the clock. On the compass, circular data can be visualized as angles, either in degrees or radians, starting at a specific point, this is our ”zero”. On the clock, we can measure specific times in terms of the 24 hours of the day. Both types can rotate either clockwise or anti-clockwise. Since we work in the unit circle, directions are just unit vectors, this simplifies its representation, making it possible to allocate points around the circle, this means there is no need take into account magnitude. The numerical representation of a given direction depends on the choice of zero and direction of rotation.

In circular statistics, circular data is used to perform statistical analysis and inference. Observations are located on the unitary circle and these angles can be represented in either degrees or radians. Another type of data that can be used in circular statistics is axial data or representations of axes, which are observations on the unit circle, any angle $\theta = \theta + 180^{\circ}$, that means there is equivalence between one observation and another one in the exact opposite direction. One way to treat axial data is to do a transformation, from θ to 2θ and use the full circle.

In the statistical background, there are a great amount of documentation on obtaining meaningful

results, in order to find how variables are affected by other variables for linear models, we know that the simplest case of a linear regression for two linear variables (Rencher, 2008) is of the form

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

where the y is the response variable, β_0 is the intercept, β_1 is a coefficient or slope of the regression and ε is the error.

Simple linear regression cannot be used on circular data directly, but the theory of linear models and its applications have been adapted to circular statistics in order to obtain useful information and make inference about a specific data set.

In circular statistics, when we want to obtain information about the existence of relationship between two variables, namely the response and explanatory variables, we can define a circular correlation and a circular regression (Lee, 2010). Correlation is a descriptive statistic about the relationship between the variables. The circular correlation has a value of zero if the variables have no linear relationship, and takes on values in the interval $[-1,1]$ generally. For the circular regression, we have the circular-linear models and the circular-circular models. The circular-linear regression is based on an infinite cylinder. For the circular-circular regression, we can intuitively describe two associations related in regards to the variables, a positive and a negative toroidal associations. The first one means that when one circular variable proceeds clockwise around the circle so does the other one, and the second one happens when we have a anticlockwise progression of one variable, so the other variable has the same sense of rotation.

When there are violations of distribution assumptions, we need to use the bootstrap technique (Efron, 1993). The bootstrap is a data-and-computer based simulation method for statistical inference, assigning measures of accuracy to statistical estimates. Bootstrap samples are generated from an original data set. Each bootstrap sample has n elements, generated by sampling with replacement n times from the original data set. Bootstrap replicates are obtained by calculating the value of the statistic on each bootstrap sample, these replicates contain information that can be used to make inferences about our data.

Some advantages about the bootstrap is that is a simple technique, though having estimators with certain complexity, we can obtain both their standard errors and confidence intervals. When having limited resources/data, bootstrap is excellent. It also good for comparison purposes for stability of our parametric results and it has asymptotic accuracy.

The simplest application of bootstrap is the bootstrap percentile method, which consists of creating a large number of replicates of a sample statistic. Subsequently we remove a small portion from the upper and lower part of the data, and the extremes of the remaining data define the confidence limits of the population. In general, resampling methods let us perform the estimation of population variables by resampling continuously, they can be used for a large number of situations.

In the following graphs we present two examples of randomly generated data sets using a uniform distribution. We generated 30 observations in the first one we generate 30 observations and 360 observations in the second one. When n=30, the observations on the circumference do not look evenly distributed, we intuitively see that 30 is a small number to obtain those results despite of the fact that the data was generated with a uniform distribution. When $n = 360$, there are almost no ”empty spots” in the circle. Observations in the second graph appear uniformly distributed.

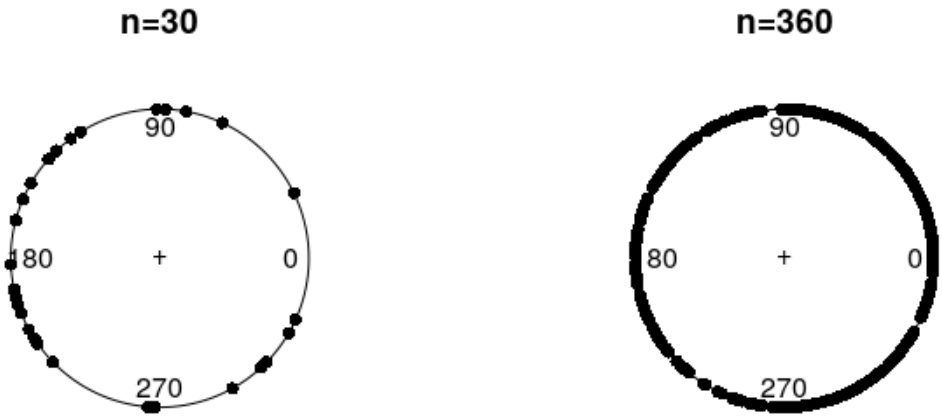

Figure 1.1: Randomly generated data with a uniform distribution for n observations

Another example of circular data is the wind directions for the months of January to June of 2011 at the Nueces Bay in Corpus Christi, Texas (See Figure 1.2.). We observe that the data are clustered towards the southeast.

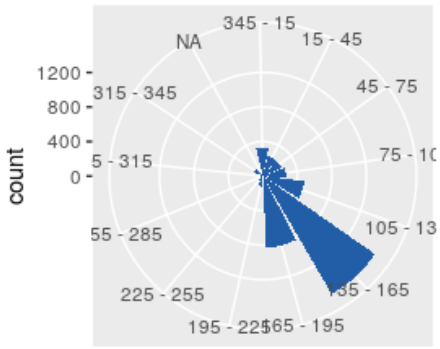

Figure 1.2: Wind rose plot of the directions for the months of January to June of 2011 at the Nueces Bay in Corpus Christi, Texas

1.2 Outline of Research

CHAPTER II: CIRCULAR STATISTICS

In this chapter we present the basic concepts about circular statistics. We emphasize the differences for the calculation of something as simple as the mean, variance, which are not the same for the linear statistics.

CHAPTER III: METHODOLOGY

We present the methods that will be used in our research. In particular, we introduce the likelihood ratio test, its assumptions, and discuss why it is useful. The bootstrap method is also in

this chapter, how this works, and how its is applied to regressions.

CHAPTER IV: DATA ANALYSIS AND APPLICATIONS

Having a big four-year data set can sometimes be complicated. To simplify, we visualize our data by different time periods. We then proceed to test our assumptions by performing different tests with different hypothesis. We also perform some regressions with different variables serving as dependent and independent, where we have the circular-circular case and the circular-linear case.

CHAPTER V: FINDINGS/RESULTS

After applying all the methods, doing some comparison and analyzing, we present our refined results.

CHAPTER VI: CONCLUSIONS

Finally, we state the conclusions for this research, and what we can do in the future in this still rich source for innovation in the field of circular statistics.

CHAPTER II: CIRCULAR STATISTICS

In this chapter we present some important concepts about circular statistics and the notation to be used during this research document.

Mean Direction

The way to properly calculate the mean for circular data (Fisher, 1993) is as follows

$$C = \sum_{i=1}^{n} cos\theta_i, S = \sum_{i=1}^{n} sin\theta_i, R^2 = C^2 + S^2 (R \geq 0)$$

where θ_i for $i = 1, 2, ...n$ are circular observations as angles.

For these unitary vectors, we get the resultant vector at the component level, so we have the following

$$cos\bar{\theta} = \frac{C}{R}, \quad sin\bar{\theta} = \frac{S}{R}$$

or

$$\bar{\theta} = \begin{cases} tan^{-1}(S/C) & S > 0, C > 0 \\ tan^{-1}(S/C) + \pi & C < 0 \\ tan^{-1}(S/C) + 2\pi & S < 0, C > 0 \end{cases}$$

Estimating the reference direction of the mode cannot be done by the arithmetic mean of all the directions. We observe in the following graph two directions 3° and 357°, clearly the arithmetic mean is 180° but both directions are close on the unit circle, this is a simple counterexample of why calculating the the arithmetic mean for directions will not work as with linear data.

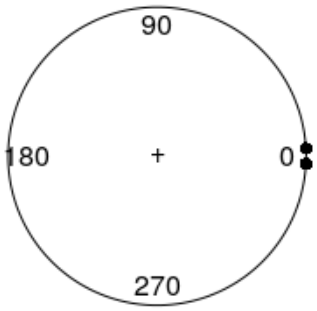

Figure 2.3: Two directions on the unit circle, 3° and 357°

Variance and standard deviation

According to (Fisher, 1993), the mean resultant length is $\bar{R} = \frac{R}{n} \in (0,1)$, the sample circular variance is $V = 1 - \bar{R}$, where $0 \leq V \leq 1$, and the sample circular standard deviation is $v = [-2log(1-V)]^{\frac{1}{2}}$.

2.1 Von Mises Distribution

A remarkable fact about the von Mises distribution is that this is the circular distribution equivalent to the Gaussian distribution on the line, its probability density function (Fisher, 1993) is

$$f(\theta) = [2\pi I_0(\kappa)]^{-1} e^{\kappa cos(\theta-\mu)} \tag{2.2}$$

with $0 \leq \theta < 2\pi, 0 \leq \kappa < \infty$, where

$$I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} e^{\kappa cos(\phi-\mu)} d\phi \tag{2.3}$$

The cumulative distribution function is

$$F(\theta) = [2\pi I_0(\kappa)]^{-1} \int_0^{\theta} e^{\kappa cos(\phi-\mu)} d\phi$$

In the following graph we have von Mises densities with different values for κ

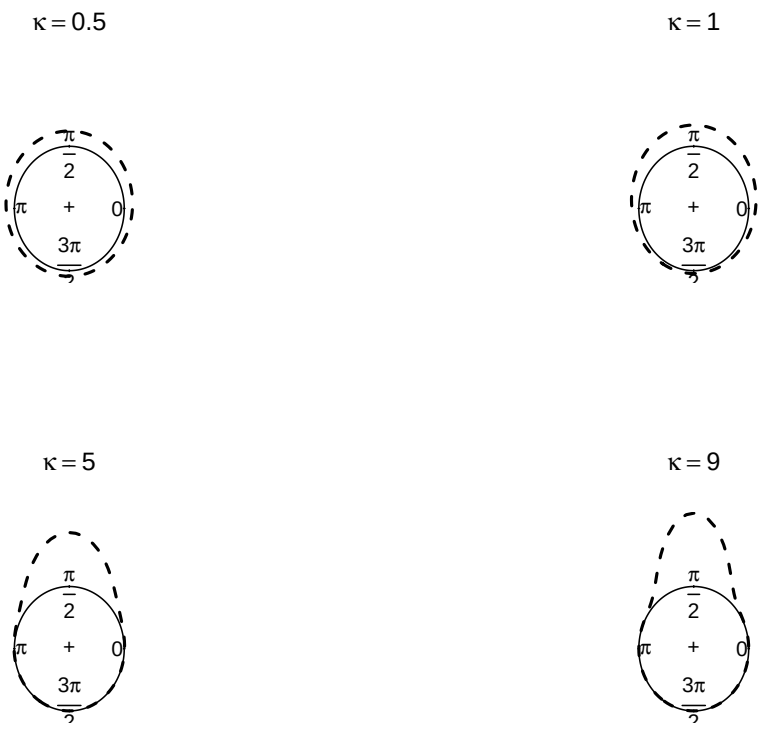

Figure 2.4: Von Mises densities

2.2 Uniform Distribution

The probability density function for the uniform distribution (Fisher, 1993) is

$$f(\theta) = \frac{1}{2\pi}, 0 \leq \theta < 2\pi$$

The cumulative distribution function is

$$F(\theta) = \frac{\theta}{2\pi} \tag{2.4}$$

with $0 \leq \theta < 2\pi$

In the following graph we have the uniform density

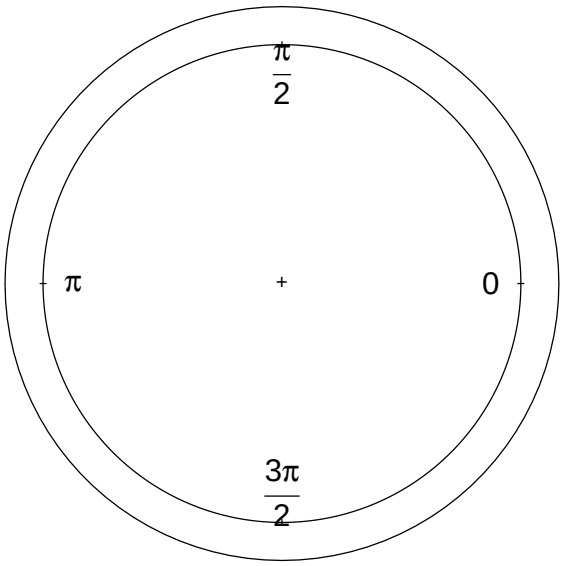

Figure 2.5: Uniform density

2.3 Circular Beta Distribution

The probability density function of the circular beta distribution (Lai, 1994) is

$$f(\theta) = \frac{1}{2^{\alpha+\beta}B(\alpha,\beta)}(1+cos(\theta))^{\alpha-\frac{1}{2}}(1+cos(\theta))^{\beta-\frac{1}{2}} \tag{2.5}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\alpha)}{\Gamma(\alpha+\beta)}$ is the Beta function and $\alpha,\beta > 0$ and $\theta \in [0,2\pi]$

2.4 The Cardioid Distribution

The cardioid distribution is also called cosine distribution. It is unimodal and symmetric, with probability density function

$$f(\theta) = \frac{1}{2\pi}[1 + 2\rho cos(\theta - \mu)] \tag{2.6}$$

with $0 \leq \theta < 2\pi, 0 \leq \rho \leq \frac{1}{2}$, where ρ is the concentration parameter.

The cumulative distribution function (Fisher, 1993) is

$$F(\theta) = (\frac{\rho}{\pi})sin(\theta - \mu) + \frac{\theta}{2\pi}, 0 \leq \theta \leq 2\pi$$

In the following graph we have cardioid densities for different values of ρ

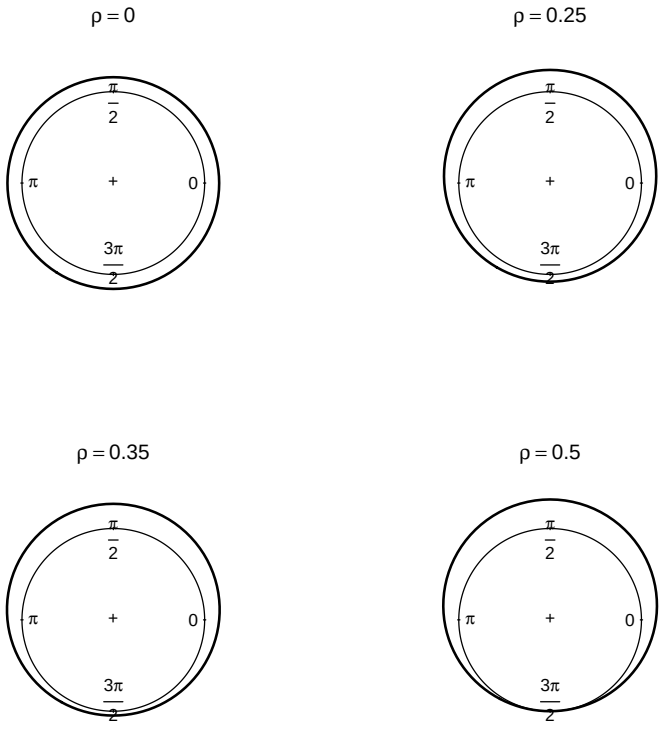

Figure 2.6: Cardioid densities

2.5 The Semicircular Normal Distribution

The probability density function of the semicircular normal distribution (Guardiola, 2006) is

$$f(\theta) = \frac{1}{\sqrt{2\pi\phi}} sec^2(\theta) exp[-\frac{tan^2(\theta)}{2\phi^2}], -\frac{\pi}{2} < \theta < \frac{\pi}{2}, \phi \in \Re^+$$

This distribution works for axial data.

The cumulative distribution function is

$$F(\alpha) = \int_{\mu-\frac{\pi}{2}}^{\alpha} \frac{1}{\sqrt{2\pi\phi}} sec^2(\theta-\mu) exp[-\frac{tan^2(\theta-\mu)}{2\phi^2}] d\theta = \frac{1}{2} erf[\frac{tan(\alpha-\mu)}{\sqrt{2}\phi}]$$

CHAPTER III: METHODOLOGY

3.1 Tests for Uniformity

We have defined the objective of our research to devise a parametric hypothesis test that can detect departures from uniformity. Currently there exist methods that can be applied to solve problems of the same nature such as the Rayleigh test and the omnibus test. We are using as the target unimodal distribution for a hypothesis test. Possible candidate distributions include the cardioid, the semicircular normal distribution, von Mises, the circular beta distribution or other plausible options.

Fulfilling the goal of obtaining a parametric hypothesis test procedure, we employ the cardioid distribution and derive a likelihood ratio test procedure. Without loss of generality, we assume that the preferred direction or seasonality is already known, then set the null hypothesis as the concentration parameter ρ equal to zero, and the alternative as ρ not equal to zero. Next we obtain both likelihood functions under the null-joint distribution and the corresponding likelihood function under the alternative hypothesis, then we are able to obtain the corresponding ratio Λ, details are in 3.2.

3.2 Rayleigh Test

The Rayleigh test (Mardia, 2000) is a test of uniformity with a von Mises alternative. We will use $u = (\lambda cos(\theta), \lambda sin(\theta))^T$ as the parameter for the distribution, where λ is the concentration parameter. The likelihood function for circular observations $\phi_1 ..., \phi_n$ is

$$l(u; \phi_1 ..., \phi_n) = nu^T \bar{x} - n \log I_0(\lambda) \tag{3.7}$$

where

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n}(cos(\phi_k), sin(\phi_k))^T \tag{3.8}$$

is the mean. Taking the partial derivative with respect to u^T, we obtain

$$W = \frac{\partial l(u;\phi_1...,\phi_n)}{\partial u^T} = n\bar{x} - nB(\lambda)(cos(\theta), sin(\theta))^T \tag{3.9}$$

When $\lambda = 0$, we have $W = n\bar{x}$ and we get

$$var(W) = \frac{1}{2n}I_2 \tag{3.10}$$

and so we obtain the statistic for the Rayleigh test

$$W^T var(W)^{-1} = 2n\bar{R}^2 \tag{3.11}$$

Theory says that for a uniform asymptotic distribution of a large sample $2n\bar{R}^2$ (Mardia, 2000) will behave

$$2n\bar{R}^2 \longrightarrow \chi_2^2 \tag{3.12}$$

3.3 Kuiper's Test

Research in circular statistics has a natural inspiration from linear models. Kuiper's test is not the exception, similar to tests in linear models obtained by comparing the hypothesized and the empirical cumulative distribution function. Kuiper's test (Mardia, 2000) is obtained in a similar way using the both cumulative distribution functions in the unit circle.

For the empirical cumulative distribution function K_n, the $\phi_1, ...\phi n$ observations become $\phi_0, ...\phi_{n+1}$, where $\phi_0 = 0$ and $\phi_{n+1} = 2\pi$, having defined the orientation and our "zero" or initial point we define K_n as

$$K_n = \frac{k}{n}, \tag{3.13}$$

where $k = 1, ...n$ and $\phi_k < \phi < \phi_{k+1}$

$$J_n^+ = \sup_{\phi}(K_n(\phi) - F(\phi)) \tag{3.14}$$

and

$$J_n^- = \sup_{\phi}(F(\phi) - K_n(\phi)) \tag{3.15}$$

where $F(\phi) = \frac{\phi}{2\pi}$ is the cumulative distribution function of the uniform distribution and K_n the empirical distribution. Finally we have Kuiper's test statistic

$$V_n = J_n^+ + J_n^- \tag{3.16}$$

J_n^+ and J_n^- are dependent on the "zero" or initial direction, but V_n does not depend on the initial direction. We consider a null hypothesis of uniformity. Kuiper's test works for uniformity against all alternatives. It is difficult to obtain an analytical form for the null distribution. The commonly used critical values for

$$V_n^* = n^{\frac{1}{2}} V_n (1 + \frac{0.155}{\sqrt{n}} + \frac{0.24}{n}) \tag{3.17}$$

are given in the table below:

α	0.1	0.05	0.025	0.01
V_n^*	1.62	1.747	1.862	2.001

3.4 Likelihood Ratio Test

The likelihood ratio test (Miller, 2014) is a method for establishing test procedures, that often produces tests with satisfactory properties. Let Ω be the parameter space, and let ω be a region within Ω. Likelihood ratio tests are defined as follows:
Define L_0 to be the maximum value of the likelihood function for $\theta \in \omega$, and L to be the maximum over $\theta \in \Omega$. Then the likelihood ratio

$$\lambda = max\frac{L_0}{L}$$

therefore we get the critical region

$$\lambda \leq k$$

where $0 < k < 1$, defines a likelihood ratio test of the null hypothesis $\theta \in \omega$ against the alternative hypothesis $\theta \in \omega^{'}$.

A very well-known result in statistics is that the likelihood ratio test $-2log\Lambda \rightarrow \chi^2_{(1)}$ under regularity conditions, this comes from (Wilks, 1938) when he proposed the test

$$W = -2log\Lambda \tag{3.18}$$

with $0 \leq W < \infty$.

The likelihood ratio test with the uniform circular distribution as the null hypothesis and the cardioid distribution as the alternative has hypotheses:

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

The likelihood function under the null-joint distribution is

$$L(\rho = 0) = \prod_{i=1}^{n}(\frac{1}{2\pi}) = (\frac{1}{2\pi})^n$$

while the likelihood function under the alternative

$$L(\rho > 0) = \prod_{i=1}^{n}\frac{1}{2\pi}[1 + 2\rho cos(\theta_i - \mu)]$$

$$= (\frac{1}{2\pi})^n \prod_{i=1}^{n}[1 + 2\rho cos(\theta_i - \mu)]$$

The ratio of two likelihood functions is

$$\Lambda = \frac{L(\rho = 0)}{L(\rho > 0)}$$

Estimation of concentration parameter P

$$\hat{\rho} = \bar{R}_1 = \frac{1}{n}\sum cos(\alpha_i) = \frac{1}{n}\sum cos(\theta_i)$$

Substituting into the likelihood ratio for ρ we have

$$\Lambda = \prod_{i=1}^{n}[1 + \frac{2}{n}(\sum_{j=1}^{n} cos(\theta_j))cos(\theta_i)]^{-1}.$$

The likelihood ratio test statistic is $-2log\Lambda$. According to (Self, 2012), the asymptotic distribution of the likelihood ratio test is a 50:50 mixture of two χ^2 random variables χ_0^2 and χ_1^2.

3.5 Bootstrap for Circular Regression Coefficients

According to (Fisher, 1992), a circular-linear regression model is

$$E(y) = \mu + 2tan^{-1}(\beta x), \tag{3.19}$$

where y is the response variable and x is the independent variable. (Fisher, 1992). We will use the bootstrap to test this model because want the verify the stability of the results we obtain from the coefficients.

There are two different approaches for bootstrapping regressions, assuming predictors are random or fixed. The first ones change in every sample and the second ones do not.

The algorithm for the bootstrap is:

- Simulation of uniform random integers with replacement, u_i for $i = 1...n$
- Bootstrap sample $s_b = (y_{ui}, x_{ui})$, $b = 1, ...B$
- Obtaining β_b using the bootstrap sample
- Find a $(1-\alpha)$ bootstrap confidence interval for the regression coefficient. When zero is in the confidence interval, we fail to reject the H_0 and if it does not contain zero, we reject the null hypothesis.

CHAPTER IV: DATA ANALYSIS AND APPLICATIONS

The data that we have used for our analysis consists of 35,064 readings taken every hour at :00 minutes, for 4 years, from January 1, 2011 to December 31, 2014. Each reading consists of wind direction and wind speed at a station in the Nueces Bay in Corpus Christi, Texas. The data has two different variables, wind direction and wind speed. The rotation direction is clockwise and the ”zero” in located in the north, wind direction measurements are in [0°,360°], while wind speed measurements were in m/s. Later on we use the first one as a the response variable and wind speed as the explanatory variable. Whenever we encountered missing data, where we lack one of the measurements, we discarded the other measurement for the same time. Our main goal is to come up with a new test for uniformity, this data will help us verify the validity of our test.

4.1 Data Analysis

In figure 4.7, we can see that for the year 2011, the months of March, May, June, July and August have a similar trend for the directions, in this case southeast. For other months, we cannot see a very well defined direction, like on January, September and December. Second, for the year 2012, the months of March, April, May, July and October have a similar trend for the directions, in this case southeast. For other months, we cannot see a very well defined direction, like on January, February, November and December. In figure 4.8, for the year 2013, the months of May, June and October have a similar trend for the directions, in this case southeast. For other months, we cannot see a very well defined direction, like on January, February, September, November and December. Finally, for the year 2014, the months of May, June and July have a similar trend for the directions, in this case southeast. For other months, we cannot see a very well defined direction, like on January, February, September, October, November and December. In general, wind directions seem to be similar though not exactly the same month by month,year by year.

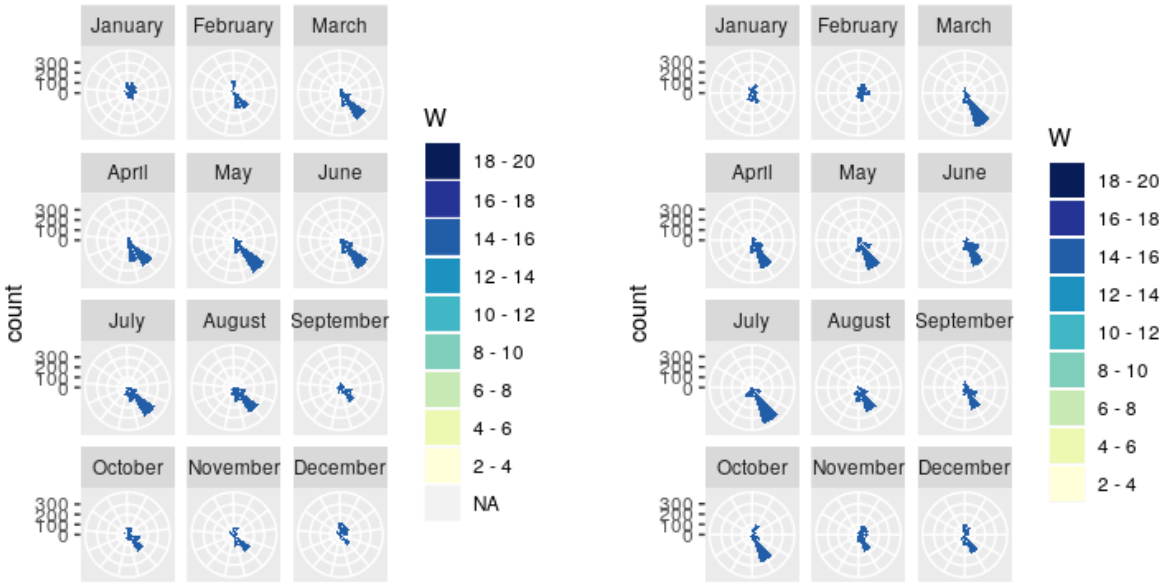

Figure 4.7: Monthly data, year 2011 and 2012

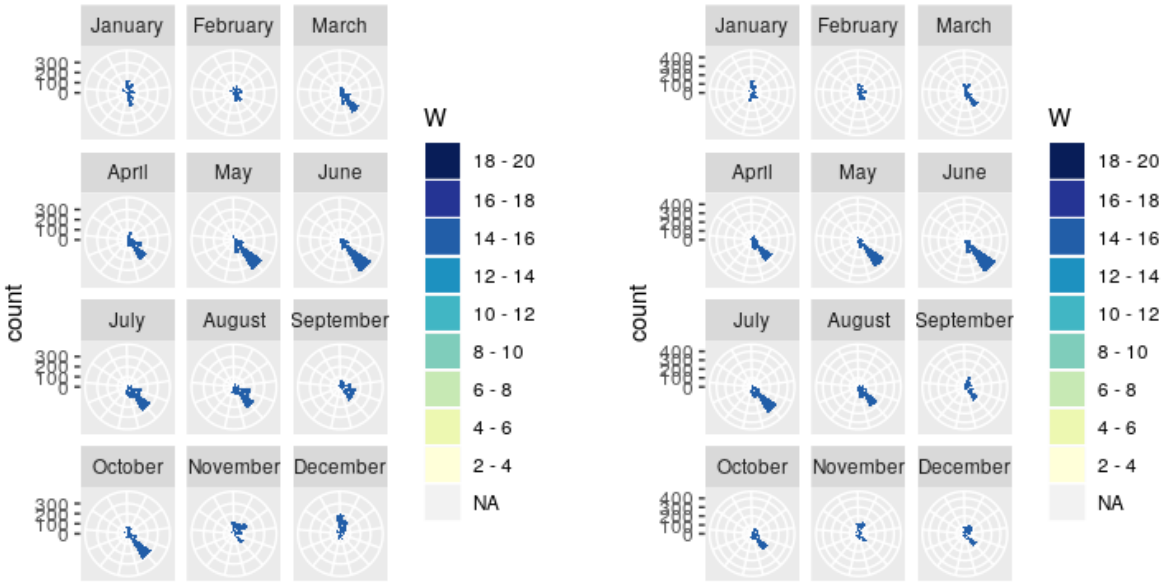

Figure 4.8: Monthly data, year 2013 and 2014

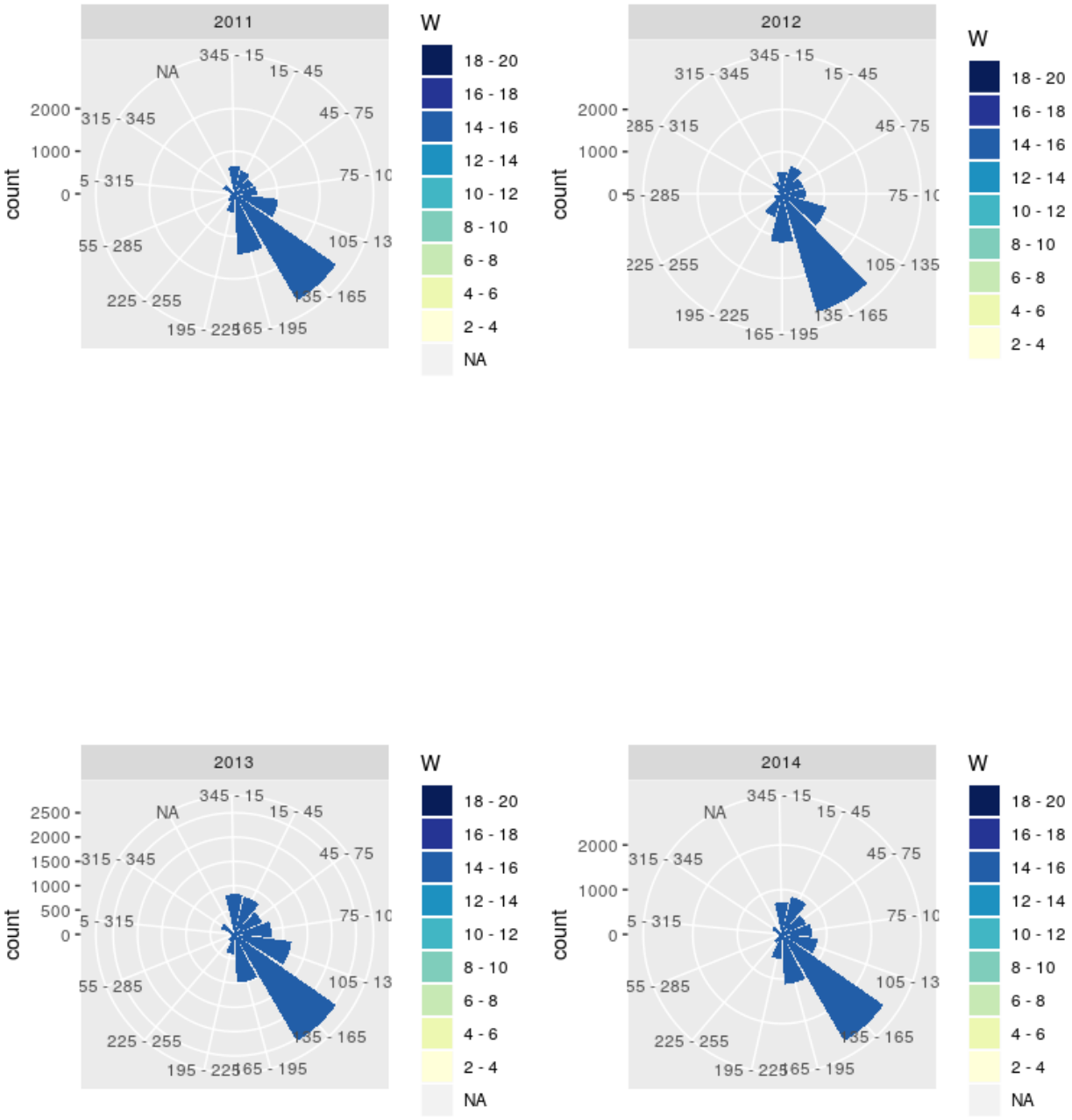

Figure 4.9: Four year data

In the following table we have all the mean directions in a monthly presentation for all the four-year data. In a general way we observe that the mean direction is consistent for every month year by year with some exceptions like February of 2011, January of 2012 and December of 2012 just to mention some.

Mean directions

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	70.69	144.77	137.51	152.41	145.47	148.4	160.05	157.97	145.1	117.87	138.4	63.04
2012	129.82	87.49	152.48	142.51	140.12	142.72	161.58	157.98	129.14	124.07	108.2	124.23
2013	75.36	96.05	131.95	121.52	144.79	148.35	149.95	105.67	104.13	128.51	71.64	94.12
2014	78.74	85.46	105.12	136.43	154.73	156.81	153.92	149.82	104.13	128.51	71.64	78.48

Four-year data analysis summary

We perform the Rayleigh test and Kuiper's test on the data. For the Rayleigh test we observe that the values for the statistic are similar for the four years, the P-value$<$ 0.0001 for the same test and period of time, because of the significant values we reject the null hypothesis of uniformity of the data. For Kuiper's test we find that the values of the statistics are similar for the year 2011 and 2012 and also a similar value for the years 2013 and 2014, all of the P-values are significant with $<$ 0.01, therefore we reject the null hypothesis of uniformity. After performing both tests, we obtain the same conclusion for the hypothesis test.

Rayleigh Test on the four-year data

The values for the test are:

Values	2011	2012	2013	2014
Statistic	0.5	0.49	0.45	0.44
P.value	$<$0.0001	$<$0.0001	$<$0.0001	$<$0.0001

All the values for the statistic similar, all P-values are significant, so we reject the null hypothesis of uniformity of the data.

Kuiper's Test on the four-year data

The values for the test are:

Values	2011	2012	2013	2014
Statistic	36.24	34.02	30.37	28.94
P-value	<0.01	<0.01	<0.01	<0.01

All of the P-values are significant, therefore we reject the null hypothesis of uniformity.

Likelihood ratio test

For computational purposes and since data behave similarly for all four years, we used the wind data for year 2011 to perform the likelihood ratio test, for the maximum likelihood estimate of the cardioid distribution we use the function "cardioid" in the R (R Core Team, 2018) package called "VGAM" (Yee, 2018). We get the test statistic

$$\Lambda = 2(logL(\rho > 0) - logL(\rho = 0)) = 4028.133$$

The test statistics is so large that it is significant. Based on the likelihood ratio test, our conclusion is to reject the null hypothesis of uniformity. Our conclusion for the tests in consistent with the Rayleigh test, Kuiper's test, with a similar significance.

The relationship between direction and speed

Our first attempt to obtain the relationship between DIR(wind direction) and SPEED(wind speed) uses circular-linear regression. We observe in the following graphs that wind direction and speed are similar for all years

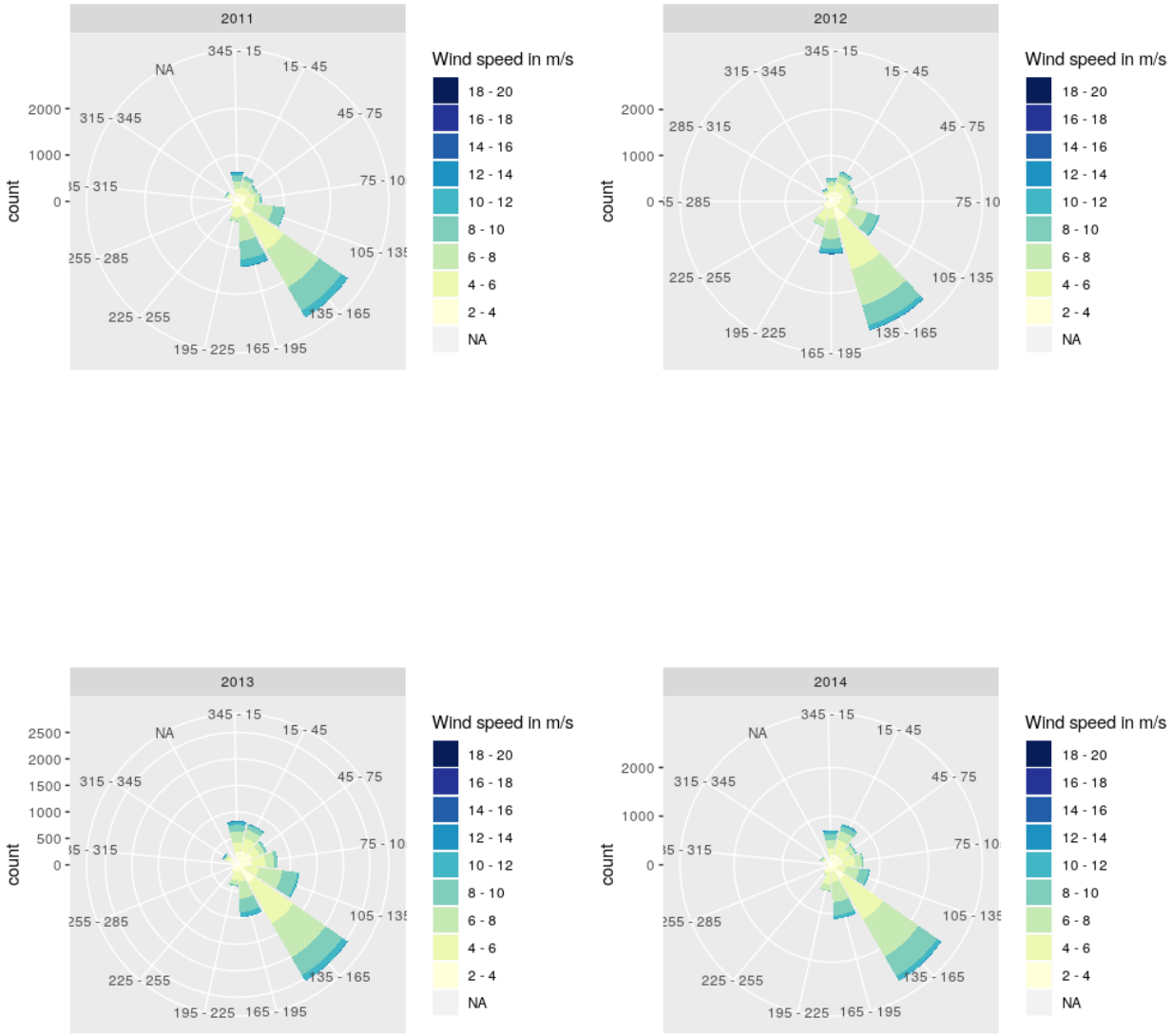

Figure 4.10: Yearly circular plots wind direction and wind speed

The circular model is given below

$$E(DIR) = \mu + 2tan^{-1}(\beta SPEED), \tag{4.20}$$

where DIR is the circular dependent variable of the model, SPEED is the linear independent variable, β is the regression coefficient. Both β and μ are unknown parameters. We assume that the response has a von Mises distribution, which is the equivalent to the gaussian distribution in circular statistics. The independent variable SPEED has the wind speed measurement corresponding to each measurement for each wind direction contained in the circular dependent variable DIR at each specific time. Equation(4.20) was originally proposed by (Fisher, 1992). We have used the open source statistical programming language R (R Core Team, 2018) to run the circular regression to estimate this relationship. We run the regression for each year for 2011, 2012, 2013 and 2014, using a data set for each year. The library "Circular" (Agostinelli & Lund, 2017) is used, the aforementioned model is included in this library with the function called "lm.circular", specifying the argument type in this case that we need a circular linear regression with type=="c-l".

The results are shown below:

Year	Estimate	Std. Error	t value	P-value	Log-likelihood	μ	κ
2011	-0.2846	0.2765	1.029	0.152	0.7569	2.421	0.01862
2012	0.1945	0.1885	1.032	0.151	0.6433	-0.3189	0.01712
2013	0.5446	0.4753	1.146	0.126	1.583	-2.357	0.02689
2014	-0.2142	0.223	0.961	0.168	0.5641	0.8173	0.01606

The standard error is < 0.3 for all four cases except for 2013 when it is close to 0.48. The P-values are consistent year by year being > 0.05 each time, showing all four cases not significant. In other words, wind speed is not useful for predicting wind direction.

The relationship between direction and time

We attempt to obtain the relationship between DIR and TIME, by a circular-circular regression, where DIR is wind direction and TIME is time of day. The circular model (Jammalamadaka, 2001) is given below

$$E(e^{iDIR}|TIME) = \rho(TIME)e^{i\mu(TIME)} = g_1(TIME) + ig_2(TIME), \tag{4.21}$$

where DIR is the circular dependent variable of the model, TIME is the circular independent variable. We have a joint probability density function $f(TIME, DIR)$ where $0 < TIME, DIR \leq 2\pi$. The model is conditional where $\mu(TIME)$ is the mean direction of DIR given TIME, $0 \leq \rho(TIME) \leq 1$ is the conditional concentration towards this direction.

$$E(cos(DIR)|TIME) = g_1(TIME), \tag{4.22}$$

$$E(sin(DIR)|TIME) = g_1(TIME), \tag{4.23}$$

we can estimate *DIR* by

$$\mu(TIME) = \hat{DIR} = tan^{-1}\frac{g_2(TIME)}{g_1(TIME)}, \tag{4.24}$$

where

$$g_1(TIME) \approx \sum_{k=0}^{m}(A_k cos(kTIME) + B_k sin(kTIME)) \tag{4.25}$$

$$g_2(TIME) \approx \sum_{k=0}^{m}(C_k cos(kTIME) + D_k sin(kTIME)), \tag{4.26}$$

So this is a general linear model:

$$cos(DIR) \approx \sum_{k=0}^{m}(A_k cos(kTIME) + B_k sin(kTIME)) + \varepsilon_1 \tag{4.27}$$

$$sin(DIR) \approx \sum_{k=0}^{m} (C_k cos(kTIME) + D_k sin(kTIME)) + \varepsilon_2 \tag{4.28}$$

according to (Ibrahim, 2013) $(\varepsilon_1, \varepsilon_2)$ is a vector whose errors are random and normally distributed. It has mean 0 and Σ is the variance matrix being unknown. Also according to (Ibrahim, 2013), A_k, B_k, C_k and D_k (these are given below in table 4.1) are parameters with $k = 0, ..., m$, where m is the suitable degree and both Σ and the standard errors can be estimated. The independent variable TIME has the hourly time of the day at 00 minutes corresponding to each measurement for each wind direction contained in the circular dependent variable DIR.
We use "R" to run the regression for the years 2011, 2012, 2013 and 2014, using only one data set for all four years. The library "Circular" is again used, the aforementioned model is also included in this library with the same function "lm.circular", specifying the argument type now in this case a circular circular regression with type=="c-c".
The results are shown below:

Table 4.1: Coefficient matrix and P-values

	cos(DIR)	sin(DIR)
Intercept	0.005992265	-0.0008707743
$cos(TIME)$	0.003933059	0.0157443501
$sin(TIME)$	0.008331328	-0.012337775
P-values	0.5055681	0.02068436

From the coefficient matrix, the first column are the estimated coefficients used for the prediction the $cos(DIR)$ and the second column are the estimates for the prediction of $sin(DIR)$. The the P-values are used to test the significance for the trigonometric terms being different from zero. In this case the P-value for $sin(DIR)$ is $0.02068436 < 0.05$, resulting significant, the model can successfully predict North-South but not East-West.

CHAPTER V: CONCLUSIONS AND FUTURE RESEARCH

Comparing results

Some advantages and disadvantages for the tests are:

Tests	Advantages	Disadvantages
Rayleigh Test (Mardia, 2000)	Modified is negligible when compared to the χ^2_2	May not work for sample small sizes
	Works for a von Mises alternative hypothesis	the regular Rayleigh test is used
Kuiper's test (Mardia, 2000)	Consistent against all alternatives and distribution-free	May not work for sample small sizes
Likelihood ratio test	Since it is smooth, it is designed to be powerful (Feltz, 2001)	May not work for sample small sizes
	for alternatives with smooth changes in the null pdf	

Conclusion

Having applied uniformity analysis to the four-year data such as Rayleigh test and Kuiper's test, we obtained significant results. For the regression models, we found that for each year wind speed is not useful in predicting our response (wind direction) but for the circular-circular model, we found that indeed time is useful when predicting our dependent variable. When we have unimodality, the best test is the Rayleigh test, for bi-modality and multi-modality Kuiper's test is our best option and for smooth departures from uniformity, a good test would be the likelihood ratio test based on the cardioid distribution.

Future research

According to (Feltz, 2001) the Kuiper test is based on the famous test on the line known as Kolmogorov-Smirnoff. One possible alternative that may be worth to explore in order to improve detection of departures from uniformity is to obtain the the average of the absolute value of J_n^+ and J_n^- instead of its sum.

REFERENCES

Agostinelli, C., & Lund, U. (2017). R package `circular`: Circular statistics (version 0.4-93) [Computer software manual]. CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA. Retrieved from `https://r-forge.r-project.org/projects/circular/`

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Feltz, C., & Goldin, G. A. (2001). Partition-based goodness-of-fit tests on the line and the circle. *Aust. N.Z. J. Stat.*(43), 207-220.

Fisher, N. (1993). *Statistical analysis of circular data*. Cambridge University Press.

Fisher, N., & Lee, A. (1992). Regression models for an angular response. *Biometrics*(48), 665-677.

Guardiola,J.H., Stamey,J.D., Seaman,J.W. and Elsalloukh,H. (2006). The semicircular normal distribution. *Far East Journal of Theoretical Statistics*(20), 207-216.

Ibrahim,S., Rambli,A., Hussin,A.G. and Mohamed,I. (2013). Outlier detection in a circular regression model using covratio statistic. *Communications in Statistics-Simulation and Computing*(42), 2272-2280.

Jammalamadaka, S., & SenGupta., A. (2001). *Topics in circular statistics.* World Scientific.

Lai, M. (1994). *Some results on the statistical analysis of directional data* (Unpublished master's thesis). The University of Hong Kong, China.

Lee, A. (2010). Circular data. *WIREs Comp Stat*.

Mardia, K., & Jupp, P. (2000). *Directional statistics*. Wiley.

Miller, I., & Miller, M. (2014). *John e. freund's mathematical statistics with applications*. Pearson.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rencher, A., & Schaalje, G. (2008). *Linear models in statistics*. Wiley.

Self, S., & Liang, K.-L. (2012). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*(82), 605-610.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*(9), 60-62.

Yee, T. W. (2018). VGAM: Vector generalized linear and additive models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=VGAM` (R package version 1.0-6)

APPENDIX A: R CODE

```
install.packages("circular")
library(circular)

#example
plot(fisherB9c, axes=FALSE, ticks=FALSE, zero=3*pi/2, stack=TRUE)
axis.circular(at=circular(seq(0, 2*pi-pi/2, pi/2)),
              labels=c("E", "N", "W", "S"))

mtext("Dance directions", side = 3, line = -2, outer = TRUE)



%mtext("Dance directions", side = 3, line = -25, outer = TRUE)%
```

```
#Uniform density


curve.circular(dcircularuniform, join=TRUE, xlim=c(-1.2, 1.2),
               ylim=c(-1.2, 1.2))


#Von Mises density plot

par(mfrow=c(2,2))

mu <- circular(pi/2) ;


kappa <- 0.5
a<-curve.circular(dvonmises(x, mu, kappa), join=TRUE, ylim=c(-1, 2), lwd=2, lty=2)

kappa <- 1
b<-curve.circular(dvonmises(x, mu, kappa), join=TRUE, ylim=c(-1, 2), lwd=2, lty=2)

kappa <- 5
c<-curve.circular(dvonmises(x, mu, kappa), join=TRUE, ylim=c(-1, 2), lwd=2, lty=2)

kappa <- 9
d<-curve.circular(dvonmises(x, mu, kappa), join=TRUE, ylim=c(-1, 2), lwd=2, lty=2)
```

```
#Cardioid density plot

par(mfrow=c(2,2))

mu <- circular(pi/2) ;

rho <- 0
curve.circular(dcardioid(x, mu, rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2)

rho <- 0.1
curve.circular(dcardioid(x, mu, rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2)

rho <- .3
curve.circular(dcardioid(x, mu, rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2)

rho <- .5
curve.circular(dcardioid(x, mu, rho), join=TRUE, ylim=c(-1.2, 1.2), lwd=2)
```

```
#Data:
fisherB9c #with circular format



#1. Example resulting being uniform-----
#Data:
plot(fisherB9c)



#Tests:
#Ho: data is uniform      CLAIM
#H1: data is not uniform



#a. Rayleigh test
rayleigh.test(abs(2*fisherB9c-360),mu=0) #this is axial data, so we had to transform it

#Conclusion
#We fail to reject null hypothesis.
#there is enough evidence to support the claim

#b. kuiper
```

```
kuiper.test(abs(2*fisherB9c-360))




#2. Example resulting not being uniform
#Data:
plot(fisherB5c)
plot(abs(2*fisherB5c))

#Tests:
#a. Rayleigh test
rayleigh.test(abs(2*fisherB5c-360))

#Conclusion
#We reject null hypothesis.
#there is not enough evidence to support the claim



#b. Kuiper test b
kuiper.test(abs(2*fisherB5c-360))
```