Statistical modelling of the rainfall and temperature

Khaya Mpehle

In this report, we statistically analyse the precipitation and temperature, indicators of climate, at a certain weather station in Darwin. Of particular concern in the analysis is the dependence of the two climatic variables, which we analyse in a copula framework. Copulas, often used in context of financial data, provide a framework for modelling the nature of dependence between different variables in a multivariate distribution. It is natural, then, to establish copulas as a tool of data analysis in climate statistics.

1 The data

The data in this investigation is monthly precipitation (in mm) and monthly maximum temperature (in Kelvin units K) as observed at a Darwin Airport weather station, station number 014015, from 1960 to 2014. This data set has been retrieved from the Bureau of Meteorology, Australia's national weather services body.

2 Results

For simplicity, we consider each of the climatic variables in the month of January, obtaining a sample of $N = 54$ for each of the temperature T and precipitation P. There are a few reasons for this. Firstly, we do not want to consider all monthly precipitations and temperatures in the sample as there will be internal, intra-yearly seasonal cycles. Taking only one data point from each year, at the same time of year, removes this intra-yearly seasonal component. Secondly, we choose January as this falls within Darwin's wet-season. This makes it likely to get a non-zero value for the precipitation in each year. Choosing a month of the year during the dry season could lead to many zeros in the precipitation data, perhaps leading to a degenerate statistical analysis. During the dry season, one may wish to see how temperature correlates to other climatic factors like wind speed or solar irradiation. Figure 1 visualises the January precipitation and temperature through a scatterplot, their histograms and box-and-whisker plots.

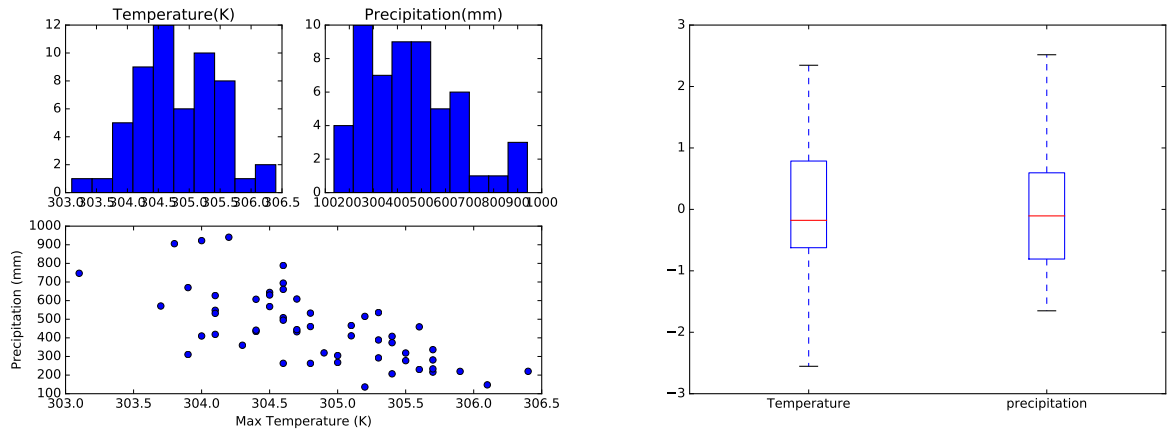

Figure 1: January Temperature and precipitation from 1960 to 2014 at Darwin airport.

Notice that the scatter plot, with temperature on the x-axis and precipitation on the y-axis, immediately indicates a negatively correlated relationship. We will measure this correlation later, but we first list some summary statistics of the data. The mean temperature across the 54 years is $\bar{T} = 304.82$ K, with a standard deviation of $\sigma_T = 0.673$ K. The mean precipitation is $\bar{P} = 452.65$ mm with a standard deviation of $\sigma_P = 192.99$ mm. To obtain the box-and-whisker plot in figure 1, we standardise each variable by subtracting their mean and dividing by their standard deviations. The box plots suggest a greater relative range in the temperature data. The spread within each each variable can be measured by their coefficient of variation, defined here as

$$c_v = \frac{\mu}{\sigma}.$$

This coefficient is a measure of signal-to-variability. As defined here, a smaller coefficient of variation means greater variability. For temperature, we have $c_v = 857.94$. For precipitation, $c_v = 2.36$. This indicates the spread in the precipitation is much greater than the spread temperature. This can be seen in the data, where we see the precipitation spread over almost an entire order of magnitude. The temperature, on the other hand, fluctuates within 2 Kelvin around 302 Kelvin. Before proceeding, we consider candidate distributions for the precipitation and temperature. The precipitation looks like it could be gamma distributed, a distribution commonly used to model rainfall precipitation probability distributions. However, notice the histogram rises to a peak, falls off, then has another small peak for large precipitation values. This histogram suggests certain large, or 'extreme', values of precipitation can occur. Heavy tail distributions display histograms in which a non-neglible number of observations occur at extreme values. For rainfall, we propose three candidate distributions: log normal, gamma and Weibull. A log normal distribution of parameters μ and σ has the density function

$$\frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right),\ x>0,$$

the gamma distribution of parameters α and β has density

$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\alpha x}}{\Gamma(\alpha)},\ x>0,$$

and the Weibull distribution of parameters k and λ has density

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k},\ x > 0.$$

While we note that the gamma distribution seems to do the best job for most of the data, the Weibull distribution does a better job on the most extreme end of the right tail. One might wish to choose the Weibull fit if care in modelling extreme events is desired, as might be the case in quantitative risk-analysis related to water damage in the wet season.

Distribution	estimates
Weibull	
k	2.53
λ	513.63
Gamma	
β	0.01198
α	5.45
Log-normal	
μ	6.02
σ^2	0.447

Table 1: Distributions fitted to the rainfall data

Figure 2 shows quantile-quantile plots of the fits of each of distribution to the precipitation data. Overall the Gamma distribution seems to provide the best fit, however notice that the Weibull distribution does well at the largest quantile. This could suggest the Weibull fit might be preferred in 'extreme event' modelling.

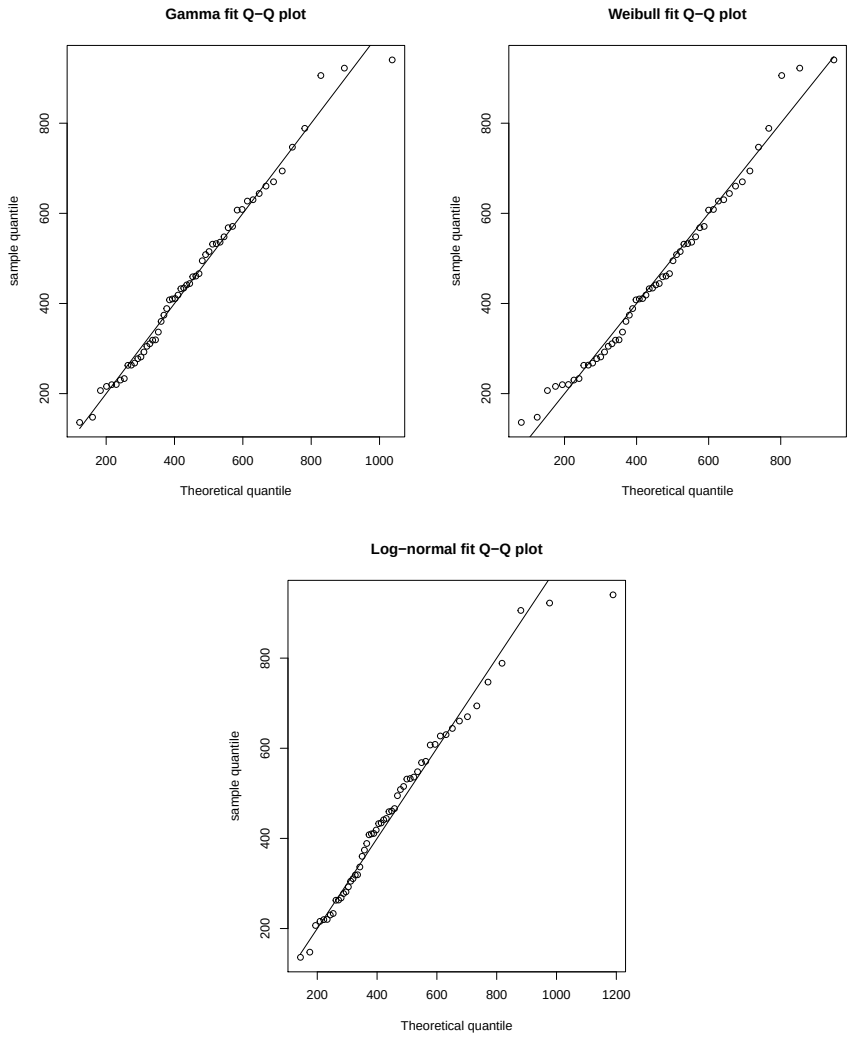

Figure 2: Q-Q plots for the three distributions fitted to the rain fall data.

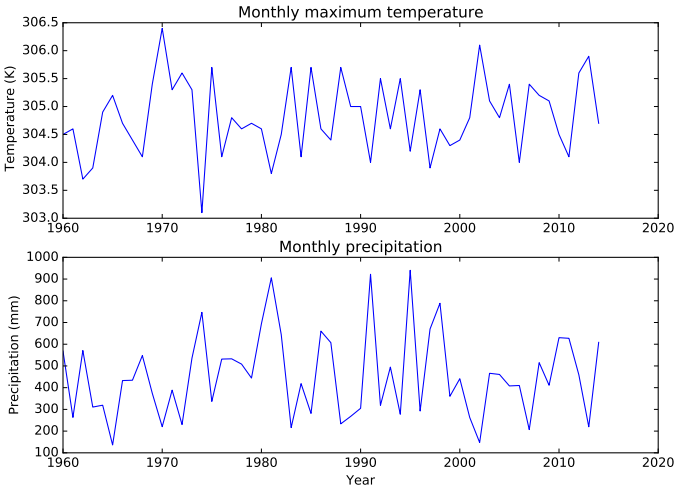

Figure 3: Time series of the January temperature and precipitation from 1960 to 2014.

Now, to utilise Copula theory, each of the temperature and precipitation, viewed as time-series, must be stationary. In figure two we see the temperature and precipitation as time-series. First we look at the auto-correlations of each time-series, as removing auto-correlations can be a crude indicator of non-stationarity [**?**]. The auto-correlation functions in figure 3 suggests that each series is weakly autocorrelated,at the 95% confidence level. An augmented Dickey-Fuller (ADF) test is implemented on each variable. For temperature, we get an ADF test statistic of -7.89, and a p-value of zero to seven significant figures. For precipitation the ADF test statistic is -7.38, and the p-value is again zero to 7 significant figures. The statistical test suggests that each time-series is stationary. With these initial tests of the data complete, we turn now to modelling their dependence.

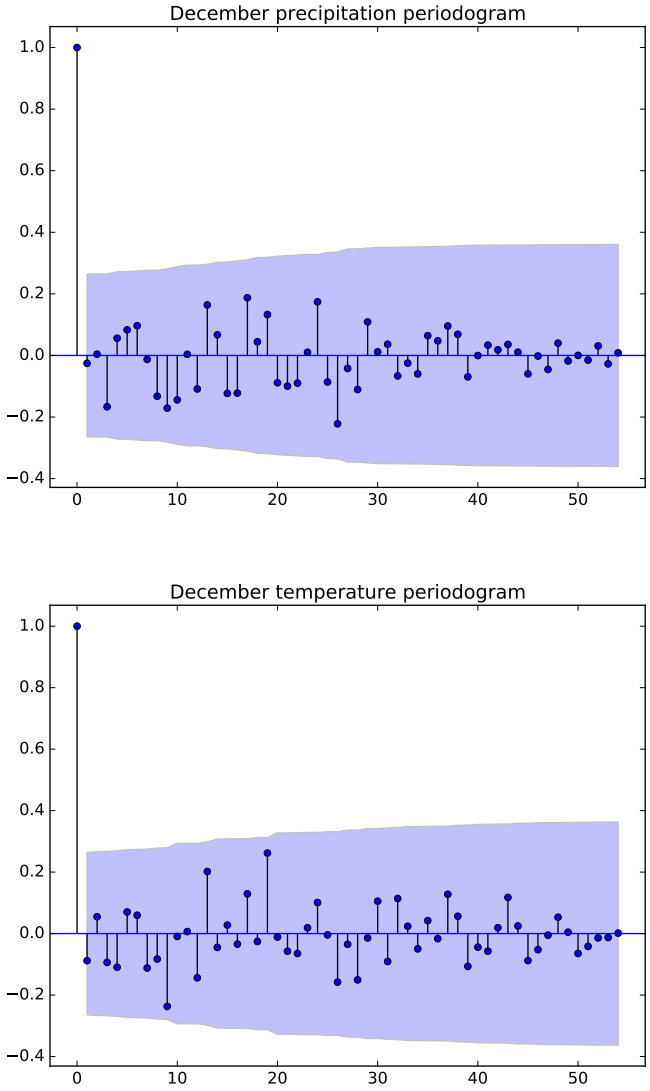

Figure 4: The autocorrelation function of the precipitation(top) and temperature(bottom). The autocorrelation function were calculated using python's statsmodels module in python. The confidence intervals are calculated with Barlett's formula.

Before estimating a copula for the data, let us first measure the correlation of the data. Specifically, we measure Kendall's correlation coefficient,

$$\hat{\rho}_\tau = \binom{N}{2}^{-1} \sum_{1 \le t < s \le N} \text{sign}((X_{t,1} - X_{s,1})(X_{t,2} - X_{s,2})), \tag{1}$$

a statistic that measures how many data points are concordant and how many are discordant. If pairs of temperature and precipitation observations are more often than not discordant, then Kendall's tau will be negative. This will be the case for a negatively-correlated relationship, which

we can easily see in figure 1. We calculate Kendall's tau to be $\hat{\rho}_\tau \approx -0.493$. While Kendall's tau gives a measure of "the degree of correlation" in the data, we wish to fit a structure to the dependence between our climatic variables using copula theory. Before proceeding, we choose to fit a copula to the bivariate distribution of the precipitation and *negative* temperature, so as to turn the negative correlation relationship into a positive correlation relationship. We choose to obtain a non-parametric fit to the copula, using the so-called pseudo-maximum likelihood fitting procedure. To do this, we first need to obtain estimates of the marginal distribution functions, and then substitute the sample into the distribution function. A non-parametric estimator of the distribution function of a random variable, X, given a sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$, is

$$\tilde{F}(x) = \frac{1}{N+1} \sum_{i=1}^{N} \mathbb{1}\left(X_i \leq x\right) \tag{2}$$

where $\mathbb{1}$ is the indicator function, equal to 1 if the condition is true and zero if not. Note that we use $N+1$ in the denomenator instead of N so that the copula estimated from the data's marginal distribution functions lies strictly inside the unit square $[0,1] \times [0,1]$. The copula sample is obtained as $(U, V) = (F_T(T_{\text{sample}}), F_P(P_{\text{sample}}))$, where T denotes temperature and the sample subscript denotes values from the sample. Otherwise, numerical problems can be encountered with copulas whose maximum likelihood functions diverge on the boundary of this unit square. The empirical distribution function for the precipitation and temperature are shown in figure 4.

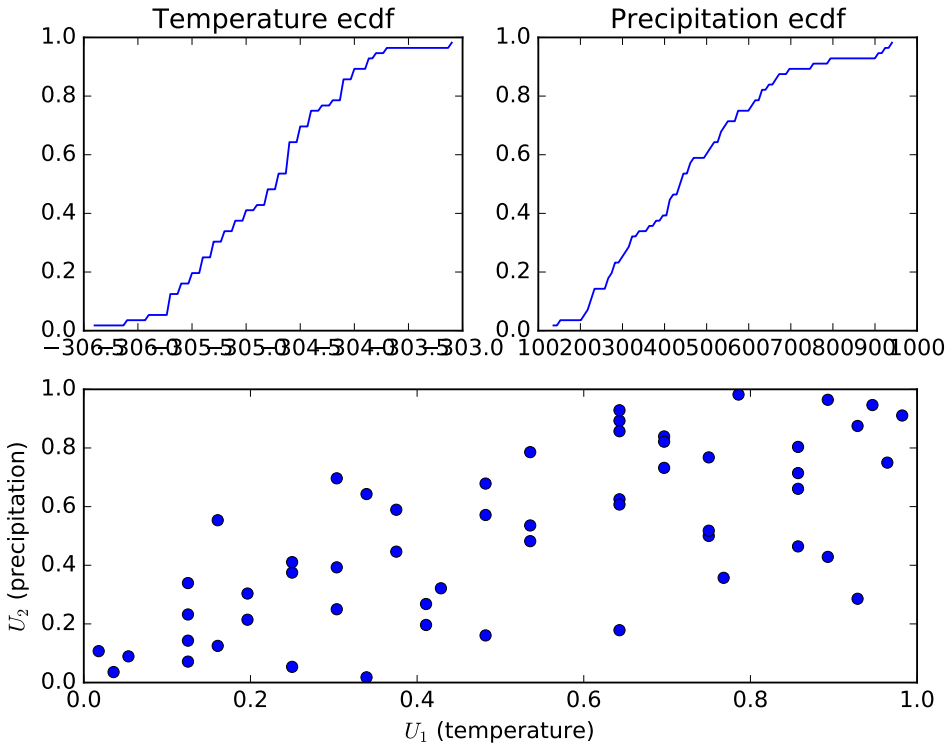

Figure 5: The empirical distribution functions of the temperature and precipitation (top) and the Copula sample obtained using these ecdfs (bottom).

Turning to the estimation of the copula, the scatter plots in figure 4 suggests that a Clayton Copula might be a decent copula, as well as Gaussian copula for reference. This is because we seem to see an asymmetric scatter in the tails of the scatterplots. Temeprature and rainfall seem to become slightly less correlated for high precipitations and low temperatures, this higher scatter visible in the top left corner of the scatter plot in figure 1. A larger sample size would be nice, but, all things considered, 1960-2014 is a good record of consistently measured temperatures. The Clayton copula is

$$C_\theta^{\mathrm{cl}}(u,v) = \left[\max\{u^{-\theta} + v^{-\theta} - 1;\, 0\}\right]^{-1/\theta} \tag{3}$$

whilst the Gaussian copula is

$$C_R^{\mathrm{G}}(u,v;\varrho) = \Phi_R\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$$
$$R = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution, and ϕ_R^{-1} is the cumulative distribution function of a multivariate Gaussian distribution with mean vector zero and covariance given by correlation matrix R. Since the correlation matrix is 2D, we have one parameter ϱ, the corelation between . The maximum likelihood procedure is implemented in R using the `copula` library. Table 1 provides a summary of the estimation of each of the two copula.

Copula	Gaussian	Clayton
Parameter	0.7178	1.523
standard error	0.063	0.276
AIC	-3.735	-3.519
S_n	0.0160	0.0378
p-value	0.1863	0.8529

Table 2: A summary of the estimated Gaussian and Clayton copulas. Note the Parameter row means the θ parameter for Clayton copula and the correlation ϱ for Gaussian copula, S_n is the Cramer-Von Mises statistic, and AIC stands for Akaike information criterion.

The parameters, standard errors on the parameters, as well as the results of a Cramer-Von mises goodness-of-fit test on the copulas are provided. The results suggest that, for this sample, the Gaussian copula is a better fit. This is owing to the Gaussian copula's smaller Cramer-Von mises test statistic and the smaller Akaike information criterion. A sample drawn from the fitted Gaussian copula is overlain on the pseudo-observations in figure 6.

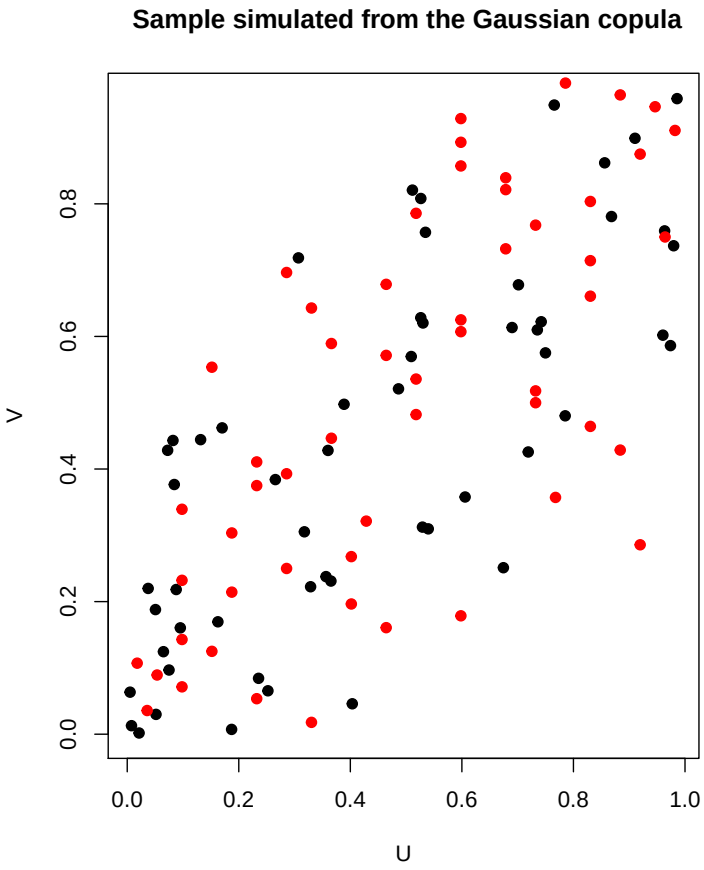

Figure 6: Pseudo-observations of the sample (red) and a sample output from the fitted Gaussian copula (black).

We summarise the report. we have investigated the dependence of rainfall and temperature at weather station [give station number], located in Darwin, NT. In general, there was a negative correlation between the temperature and rainfall during January, a month falling in the region's wet season, from 1960 to 2014. Regarding the marginal distribution of the rainfall, we suggested that while a gamma distribution is a good fit to the precipitation data, although a heavy-tailed distribution may also be an appropriate fit. Such an observation is important to make as tail-events in multivariate data can affect the copula chosen. In fitting a copula to the data negative temperature modified sample, we found the Gaussian copula to be a better fit to than the Clayton copula. Improvements to the investigation would include a large sample size in the data. It turns out the BOM has precipitation data dating back to 1900, located on a section of its website that we only found out about after we commenced this investigation. This larger data set would improve the power of the statistics performed here. We note that data dating back to 1900 may produce a temperature trend due to the effect of global warming. However, Darwin's location on a coast in the tropics may mean the trend is not detectable.