Swiss Rainfall Modelling

Harrison Zhu
EPFL, Switzerland

June 13, 2018

Summary

In this investigation, we present rainfall modelling techniques using generalised linear models (GLMs). The data is collected from 1 January, 1971 to 31 December, 2008, and is rather sparse as many days were not recorded to have rained. We modelled both the probability of having rainy day and its rainfall amounts using binomial and gamma regressions. The main source of our motivation comes from a well known paper by Stern and Coe [2], where they explored modelling the states of the days $\{\text{no rain, rain}\} \equiv E = \{0, 1\}$ using m-th order Markov chains, where $m \in \mathbb{N}$. In addition, we fitted Fourier series as the predictors for our GLMs. The binomial regression seems to fit well for the probability and the gamma regression seems to fit well for the amount. We discovered and statistically confirmed that there are obvious yearly patterns for both the probability and amount.

Contents

1 Initial Data Analysis and Treatment

1.1 Description of the Dataset

The data is collected from 1 January, 1971 to 31 December, 2008 i.e. there are 13880 observations. There are a lot of days without any rainfall, but if we count the amount of rainfall for each day of the year we can already conjecture a pattern (see Figure 1, left): that there is a peak between the days 100 and 170, which is between March and July. The question that we will ask ourselves throughout this report is whether there are days of the year when it is more likely to rain, and the converse.

On the other hand, if we model the amount of rainfall, are there days that are more rainy than others? Evidently from the figure (see Figure 1, Right) we can see an interesting dip in the mean rainfall between days 100 and 170. It seems that there are more occurrences of rain between these days but the amount of rainfall seems not to be a very significant.

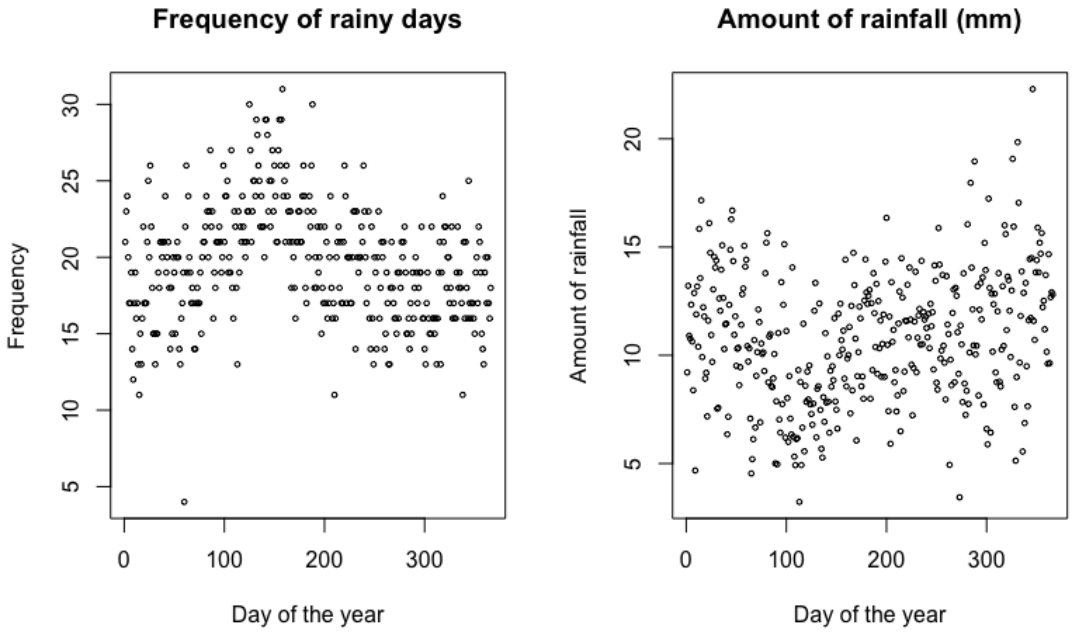

Figure 1: (Left) Frequency of rain in a year and (Right) mean amount of rainfall (mm) in a year calculated over 38 years.

1.2 Setup and Pre-processing

First of all, we impose as suggested $\xi = 0.05$mm to be some threshold so as to judge whether there was really rain or not. This is because it is often the case that anything less than this threshold will either be due to a mechanical error or amounts of rainfall that are extremely insignificant. Next, to treat the data for our models, we numbered the days each year that we are interested in from $t = t_1, .., t_T$. For $T = 366$, it would entail

numbering all the days in each year from 1 to 366, where 60 denotes the extra day, the 29th of February, during a leap year. For $T = 73$, we take the days $5i - 4$ for $i = 1, .., 73$. The models that interest us are derived from the second-order Markov chain $J(t)$, as described in [2].

Definition 1.1. Let $J = \{J(t)\}_t$ be a discrete Markov chain on the *states* {no rain, rain}$\equiv$ $E = \{0, 1\}$ of the day, where

$$J(t) := \begin{cases} 1, & \text{if day } t \text{ rains,} \\ 0, & \text{if day } t \text{ does not rain,} \end{cases}$$

for $t = t_1, .., t_T$. Furthermore, J is a second-order Markov chain if J is time homogeneous and satisfies the second-order Markov property. In particular,

$$\mathbb{P}[J(t) = 1 | J(t-1), J(t-2), ...] = \mathbb{P}[J(t) = 1 | J(t-1), J(t-2)].$$

The *transition* probability $p_{hi}(t)$ is such that

$$p_{hi}(t) := \mathbb{P}[J(t) = 1 | J(t-1) = i, J(t-2) = h],$$

where $h, i \in E$. Similarly, we can extend the definition of J to m-th order Markov chains.

Definition 1.2. The number of transitions are sufficient statistics for $p_{hi}(t)$ [2] and so define it to be

$$n_{hij}(t) := \text{Number of days with } J(t) = j, J(t-1) = i, J(t-2) = h,$$

for $h, i \in E$ and $t = t_1, .., t_T$.

The *observed* estimates of $p_{hi}(t)$ is defined to be the ratio

$$r_{hi}(t) := n_{hi1}(t) / n_{hi\cdot}(t),$$

where $\cdot$ indicates a summation over that index.

For us, $T = 365$ and $T = 73$ will be of interest to us, as in particular for $T = 73$ later we are *pooling* over 5 days. That is, we calculate the mean frequency of rainy days across 5 days and repeat for all $73 \times 5 = 365$ days, dropping the 366th day. Note that as we only have 38 years, leaving the 366th day untreated should not affect the overall models later on too much. Lastly, we also calculated the mean rainfall over all the 366 days and the pooled values over 5 days. With this we conclude our pre-processing and begin modelling.

2 Modelling

Here we present 2 approaches to extracting information from our data. We first build 3 models to model the probability of rain throughout a year, and then build 3 models for the mean rainfall amounts.

2.1 Probability of Rain

We first examine what our processed number of days $n_{hij}(t)$ looks like. From the points in Figure 3 we can see clear trends for $p_{00}(t)$ and $p_{10}(t)$ whereas for the other 2 the probability distributions seem mostly linear and the points are roughly homoskedastically spread around its mean. Note that this data can be modelled using a binomial regression

$$n_{hi1}(t) \sim Binomial(n_{hi\cdot}(t), p_{hi}(t)).$$

The choice of the link function h, such that $p_{hi}(t) = h^{-1}(g_{hi}(t))$ [2], is the standard logit function for us, and choosing the predictor $g_{hi}(t)$ will be the main focus of our investigation.

2.1.1 Models 1 and 2: Feature identification

To detect the patterns, we first deploy 2 logistic regression models without the Markov chains. For model 1, we model $\{J(t), t = 1, .., 13880\}$, the binary stochastic process on the states E, where 13880 is the number of days in our dataset, against the number of the day. Each day will have different rainfall probabilities. For model 2, we again do something similar and model the number of rainy days per day of the year against the days 1 to 366.

We can see from the results of model 1 (see Figure 2) that there is a clear seasonal trend, perhaps sinusoidal, for the probability of rainy days. Certainly from the observed proportions of rainy days and model 2 (see Figure 2) we can spot a clear sinusoidal trend, that there is a peak between days 100 and 170.

More rigorously, we suppose that for a GLM we have that the true predictor $g(t)$ is a real-valued piecewise continous function on $t \in [0, 367] \subset \mathbb{R}$. Then the Fourier series $S_m(g)$ of g with m harmonics converges pointwise to g almost everywhere as $m \to \infty$ in the normed vector space $(L^2([0, 367]), ||\cdot||)$, such that $||f|| = (\int_0^{367} |f(t)|^2 \, dt)^{1/2}$ (see [3]). Thus, we can obtain an estimate for the true predictor g for our GLMs and this suggests that we could possibly try fitting Fourier series for our non-Markov and Markov models [2].

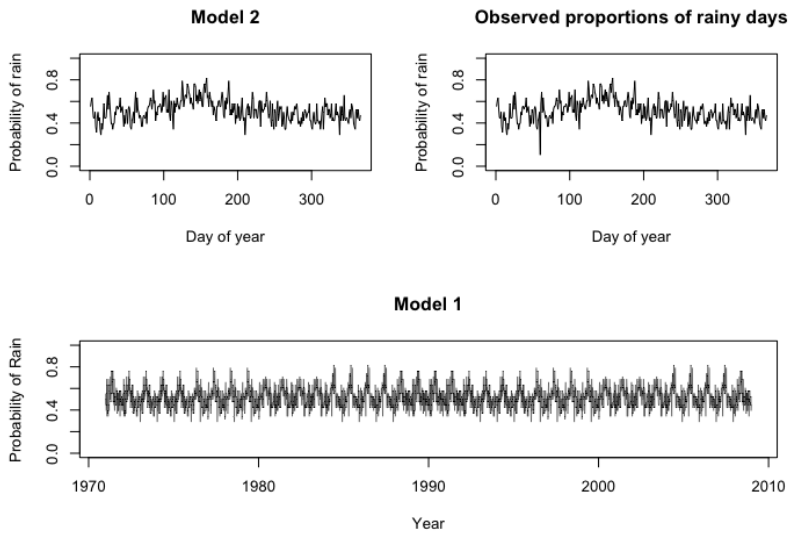

Figure 2: (Top Left) Model 2 results. (Top Right) Observed proportions of rainy days. (Bottom) Model 1 results.

2.1.2 Model 3: Non-Markov chain model

For model 3, we begin by fitting Fourier series to a simple non-Markov model. Let $N(t)$ denote the total number of days, $n(t)$ denote, similar to what we defined earlier in section 1, the number of rainy days, and $p(t)$ to be the probability of rain on day $t = t_1, .., t_T$, with $T = 366$. Then, we fit the model

$$n(t) \sim Binomial(N(t), p(t)),$$

with $logit(p(t)) = a_0 + \sum_{k=1}^{m} a_k \sin kt' + b_k \cos kt'$ and $t' = 2\pi kt/366$ [2], and $m \in \mathbb{N}$ being the number of harmonics that we would like to fit. We obtain from the below results (see Table 1) that $\boldsymbol{m = 5}$ harmonics seems suitable, after using the **forward selection algorithm**, at 5% level . That is, we begin from a simple predictor with just the intercept and apply likelihood ratio tests consecutively by adding 1 harmonic each time. We can see from the results (see Figure 3) that the Q-Q plot of the modified residuals r_j^* ([1], slide 39) is good and the points are mostly within the 95% confidence bands, although the left lower tail seems light (detached from the line to the left). Otherwise we can clearly see from the fitted probabilities that the trend has been well-estimated. The Cook's distance shows that only 3 points out of 366 have breached the red line (defined by $8/(n-p)$, where n is the number of samples and p is the number of parameters). Thus this indicates that there is a lack of outliers and influential observations. Furthermore, despite that the sequential use of likelihood ratio test may give us spurious results ([1], slide 65), since the AIC of the model is reported to be 1868, which is small compared to simpler models described on Table 1, we choose this model over other similar models.

Finally, the Pearson's statistic given by $X^2 = 354$ has a theoretical χ^2 distribution with

degrees of freedom equal to 355. Thus at 5% level, since our p-value is 0.5, this indicates that there is significant evidence to suggest that our model is well-fitted.

m	Likelihood ratio statistic $2(D_{m-1} - D_m)$	p-values	AIC
0	-	-	2039
1	56	≈ 0	1902
2	11	5.9×10^{-13}	1878
3	6.1	0.0051	1876
4	30	0.046×10^{-7}	1877
5	4.4	3.0×10^{-7}	1866
6	1.8	0.11	1868

Table 1: p-values for the sequential likelihood ratio test (to see whether we want to adopt m harmonics). The likelihood ratio statistic is defined as $2(D_A - D_B)$, where model B is nested in model A [1], and D is the residual deviance. The p-values are obtained via the χ^2_2 distribution.

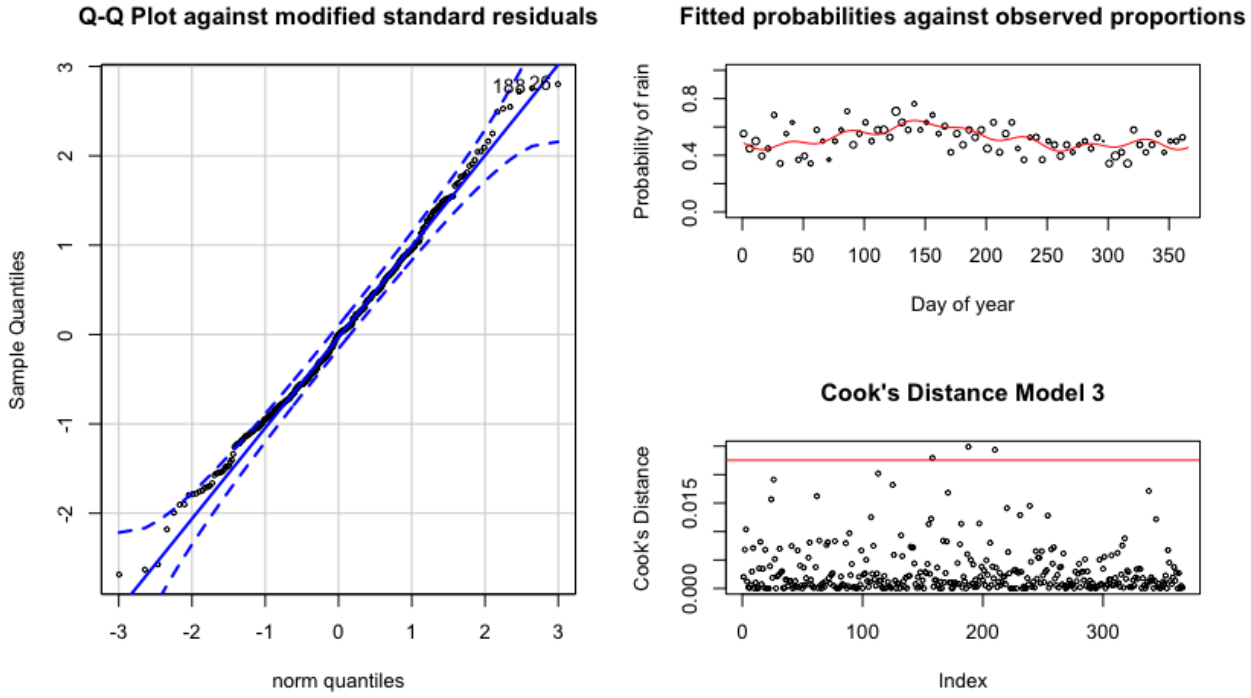

Figure 3: (Left) Plot of r_j^* against the standard normal quantiles with 95% confidence bands(blue dotted lines). (Right Top) Fitted values (red line) against observed proportions. (Right Bottom) Cook's Distance.

2.1.3 Model 4: Markov chain models

Assuming now that $J(t)$, for $t = t_1, .., t_T$ is a second-order Markov chain, we now proceed to make a fit for each of the 4 cases of $(h, i) \in E^2$. For completeness, we have displayed the residual deviances and its degrees of freedom for each case up to $m = 5$ on Table 9. Again using the forward selection algorithm, for the daily data we choose $\boldsymbol{m = 2, 0, 2, 3}$ for cases (1,1), (1,0), (0,1) and (0,0) respectively. The results are presented below (see Figure 4), and we can see that again we have successfully captured the main trend. As suggested by [2], we have made the sizes of the points of the proportions relative to the number of rainy days observed. This gives us a visualisation of whether the model has fitted well or not as we have considered the number of rainy days observed as the 'weights'.

Furthermore, by looking at the Q-Q plots (not shown) of the 4 models most of the points lie roughly on the diagonal line, indicating that the residuals r_j^* are approximately normally distributed ([1], slide 39). The Cook's distance plots (not shown) also display most of the points below the $8/(n - p)$ line, indicating a lack of outliers and influential observations. A final check on the Pearson's statistics reveals that all the p-values are way above 5%, indicating adequate fits. Again, we check the AICs of the models and they all appear to be relatively low compared to simpler and more complicated models. While we do have slightly lower AICs for more complicated models, we keep the forward-selected models for parsimony.

Finally regarding the model with 5-day pooling $T = 73$, we obtain $\boldsymbol{m = 0, 0, 1, 1}$ for (1,1), (1,0), (0,1) and (0,0) respectively. The fit is great (see Figure 5). We can see that we have captured the trend well, especially when there are many occurrences of rain. The Q-Q plots (not shown) of these 4 models also show that a few of the tail values lie outside the standard normal confidence bands.

(a) Daily (h, i)

m	d.f.	(1,1)	(1,0)	(0,1)	(0,0)
0	365	403.6	437.4	466.5	610.85
1	363	393.0	436.2	445.5	458.06
2	361	384.0	435.1	440.5	446.73
3	359	383.5	432.6	439.8	441.68
4	357	381.0	430.9	439.0	439.39
5	355	375.3	430.5	431.8	431.97

(b) 5-day pooling (h, i)

m	d.f.	(1,1)	(1,0)	(0,1)	(0,0)
0	72	19.7	28.0	24.3	37.8
1	70	17.5	27.7	19.1	18.6
2	68	15.5	27.4	18.3	17.3
3	66	14.4	26.7	17.5	16.4
4	64	14.4	26.3	16.5	14.1
5	62	14.4	24.4	15.8	13.9

Table 2: Residual deviance and degrees of curves fitted with m harmonics on a daily basis and with 5-day pooling.

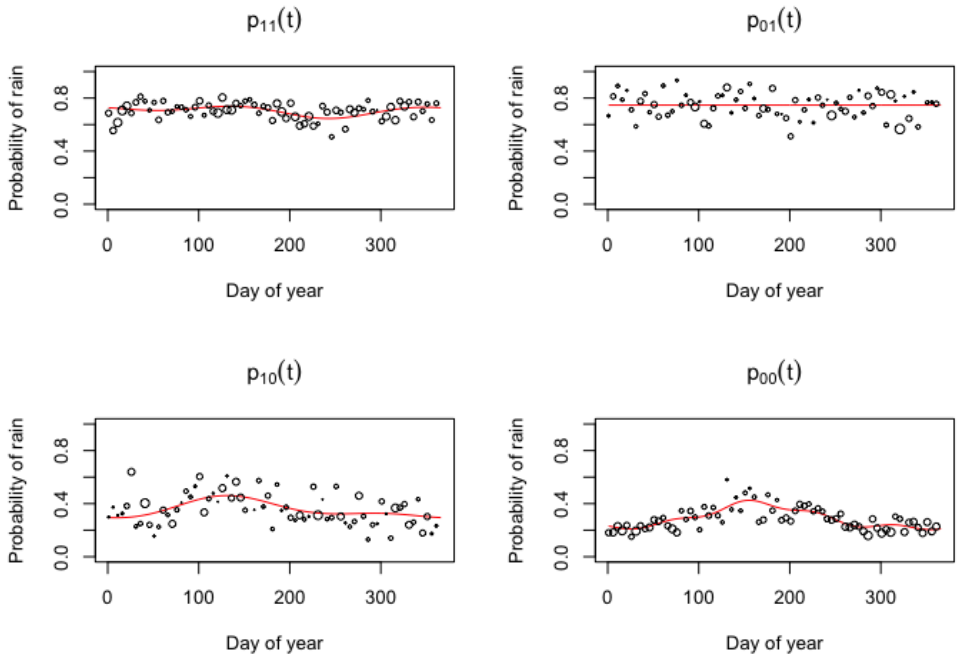

Figure 4: Plots of the fitted probabilities (red line) with the observed proportions for the daily fitting.

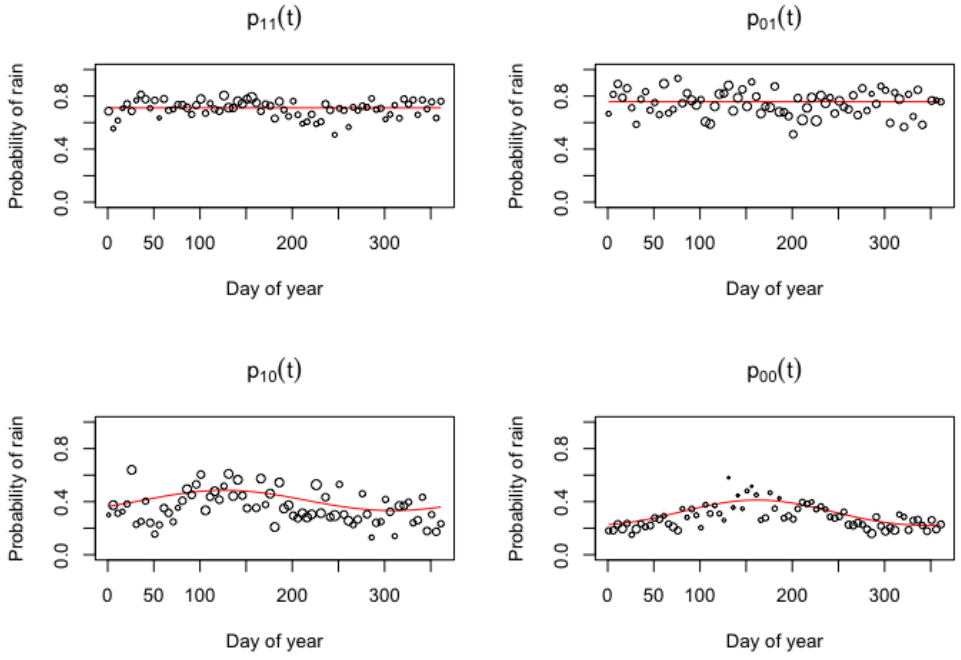

Figure 5: Plots of the fitted probabilities (red line) with the observed proportions with pooling.

2.2 Rainfall Amount

2.2.1 Non-Markov rainfall modelling

We model the amount of rainfall $X(t)$ per day conditional on $J(t) = 1$ ($J(t)$ for $t = t_1, .., t_T$ is a binary stochastic process on the states of the days) using a gamma distribution on $X'(t) = X(t) - \delta$, where we set $\delta := \min_t X(t)$, with observations $x_j(t)$, for $j = 1 \to n(t)$ and $n(t)$ corresponding to the number of years with rain on the t-th day suggested by Coe et al. [2]:

$$X'(t) \sim Gamma(\kappa, \kappa/\mu(t)),$$

where κ is the shape parameter (that we will estimate) and the scale parameter $\mu(t)$ is defined as $\log(\mu(t)) = g(t)$, with $g(t)$ being the predictor again. Again, noting that $\sum_{j=1}^{n(t)} X_j(t) = n(t)\bar{X}(t)$ is a sufficient statistic [2], we fit $\bar{X}(t)$ to a gamma distribution with predictor $g(t)$ again. Note our F statistic for the Analysis of Deviance ([1], slide 57) will correspond to

$$F = \frac{D_A - D_B}{p - q}\hat{\kappa} \dot{\sim} F(p - q, N - q),$$

where $\hat{\kappa}$ is the estimated shape parameter via the method of moments, N is the sample size, q and p are the degrees of freedom of models A and B with $q < p$. For our investigation, we estimate the shape parameter κ by first fitting a full model with $m = 5$ harmonics, and then proceed with the Analysis of Deviance with the F-statistic above.

For the daily model $T = 366$, we obtain that $\boldsymbol{m = 4}$ via the forward-selection algorithm. The AIC is also one of the lowest and so we choose this order of harmonics for our Fourier series. For the 5-day pooling model, our forward-selection algorithm and AIC criterion (lowest AIC) yields $\boldsymbol{m = 3}$. We can see above in Figure 6, top, of the daily fitting that we have clearly estimated the trend of the mean rainfall, again, especially when there is a high occurrence of days with rainfall. Similarly, the same goes for the 5-day pooling. However, the Q-Q plots (see Figure 7) of r_j^* reveal the we have heavy tail behaviour for the daily model. Pooling seems to have removed the non-normal tail behaviours. Now taking a look at the Cook's distances, for the daily fitting there is only 1 value that breaches the threshold $8/(n - p)$, but for the pooled fitting many of the points have breached that threshold. This suggests that our models have some limitations and so this leads on to adopting second-order Markov chains again.

(a) Daily $\hat{\kappa} = 0.067$

m	d.f.	Residual deviance	F statistic
0	365	30.3	-
1	363	28.7	1.64
2	361	26.8	2.09
3	359	26.0	0.55
4	357	25.8	0.22
5	355	25.8	0.06

(b) 5-day pooling $\hat{\kappa} = 0.016$

m	d.f.	Residual deviance	F statistic
0	72	1.92	-
1	70	1.59	0.33
2	68	1.17	0.42
3	66	1.06	0.11
4	64	1.02	0.04
5	64	1.02	0.01

Table 3: Analysis of Deviance table for (a) Daily mean rainfall and (b) Pooled rainfall amounts.

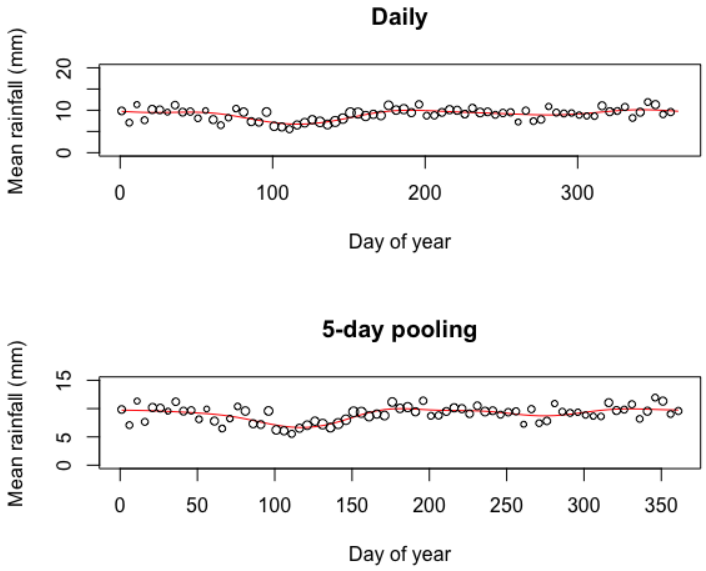

Figure 6: Plots of the fitted mean rainfall (red line) with the observed mean rainfall. (Top) Fitted with daily data. (Bottom) Fitted with 5-day pooling.

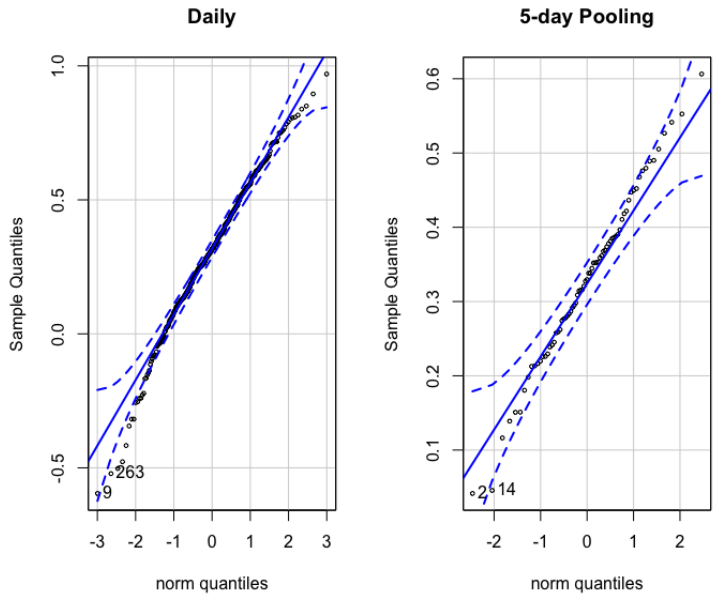

Figure 7: Plots of r_j^* against the standard normal quantiles with 95% confidence bands(blue dotted lines). (Left) Daily fitted rainfall. (Right) 5-day pooling fitted rainfall.

2.2.2 Second-order Markov chains

Similar to what we did just now, we model $X'(t)|J(t) = 1, J(t-1) = h, J(t-1) = i$, where $h, i \in E$, with $J(t)$ for $t = t_1, .., t_T$ being a second-order Markov chain on E. We obtain for the transition probabilities of $(1,1), (1,0), (0,1), (0,0)$ the harmonics $\boldsymbol{m = 4,1,1,3}$ respectively via the forward selection algorithm. Again, we also reviewed the AIC values and they suggest us to keep these harmonics that we obtained. However, when we plot the residuals r_j^* against the standard normal quantiles we have an issue with the left tails being too heavy (see Figure 8), suggesting that our models are not well-fitted.

However, if we pool the data over 5 days we are able to improve our model fit. With the pooled data, we obtain for $(1,1), (1,0), (0,1), (0,0)$ the harmonics $\boldsymbol{m = 4,1,1,2}$. The Q-Q plots (not shown) now have most of the points contained in the 95% confidence bands and along the diagonal lines. The results of our fit are shown on Figure 9. As we can see, the important features are well calculated, especially when there are more frequent occurrences of rainfall. Finally to note, we have also managed to contain most of the points in the Cook's distance plot below the threshold $8/(n-p)$, indicating that influential observations have been successfully dealt with through pooling. Therefore the evidences suggest that using these second-order Markov chains are better ways to model rainfall amounts than 0-order Markov chains.

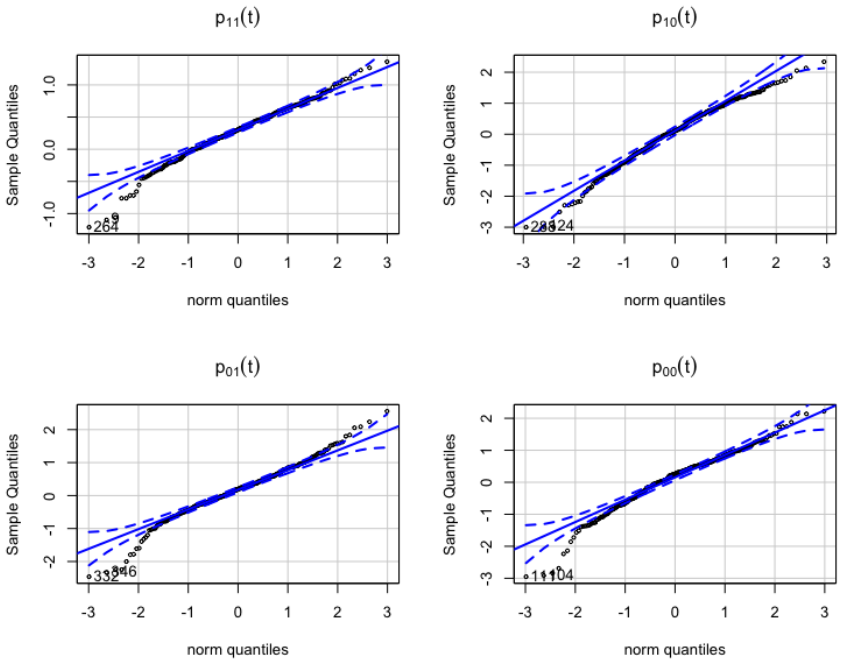

Figure 8: Plots of r_j^* against the standard normal quantiles with 95% confidence bands(blue dotted lines) for daily fitting.

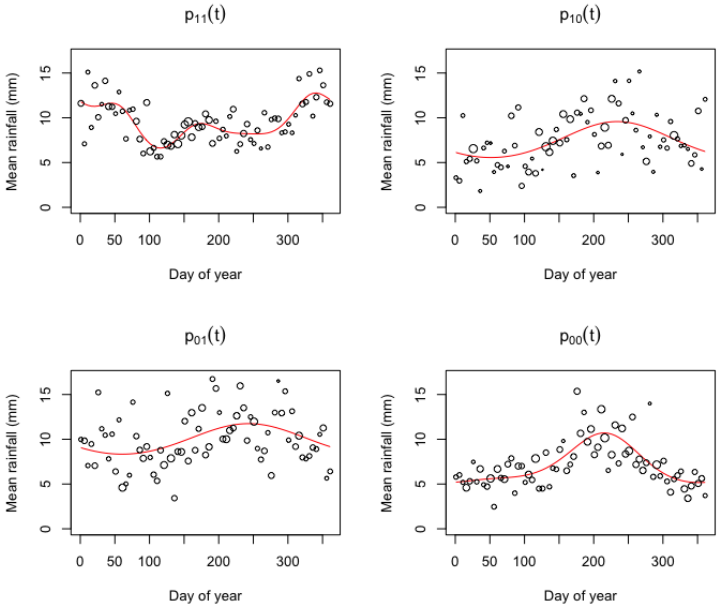

Figure 9: Plots of the fitted mean rainfall (red line) with the observed mean rainfall with 5-day pooling.

3 Conclusion

To conclude our study, we first hypothesised by examining the original data that there are obvious trends that we can spot for both the frequency of rainy days and amount of rainfall per day of the year. We have strong evidence to suggest that the data is seasonal, yearly, by doing a simple logistic regression using all the 13880 days as factor covariates. We then spotted the sinunoidal behaviours of both the proportions of rainy days and rainfall amounts and changed our predictors to Fourier of harmonics m, where m was to be determined by our study.

For the probabilities/proportions of rainy days, we obtained different results for when we assumed that the states of the days was a discrete Markov chain. Both types of models fitted well, and both give us different and relevant information about the occurrence of rainy days. For the latter, it can be practically used for live weather channels as it will give us a more accurate prediction, since we can also consider 4 cases depending on the states of the previous 2 days that we are trying to predict. For the former, it may be insightful to model what the general pattern or trend is for rainy days in a year.

For rainfall amounts, we first assumed the non-Markov structure of the states and obtained a good fit when using 5-day pooling. Similarly, when we introduced the second-order Markov chains again we also concluded that pooling had a positive effect on the fit.

During our study, we first conjectured that while the occurrence of rainfall peaks between days 100 and 170 during a year, there is actually a trough during this period for the amount of rainfall. This may seem unintuitive but this was in fact also shown by our models. Therefore, the insight that we can extract from this is that possibly despite the more frequent occurrences of rain, each rainy day actually contributes very little to the rainfall amount. This could suggest that short showers are more likely to occur during the spring-summer transition period.

References

[1] Davison, A.C., (2018). Modern Regression Methods. *EPFL*.

[2] Stern, R. D. and Coe, R. (1984) A model fitting analysis of daily rainfall data (with discussion). *Journal of the Royal Statistical Society series A* **147**, 1-34.

[3] Stein, E. M. and Shakarchi, R. (2007) Fourier Analysis: An Introduction. *Princeton Lectures in Analysis*, 1, page 69-78.