South Africa drought and wildlife survival

Andrea Corrado 20205529

Contents

Contents

Abstract

1 Introduction

During last decades, interest and awareness about the climate change has steeply increased. There have been several studies examining and reporting the potential consequences of these phenomenon. For instance, the temperature and rainfall change has been of particular interest. Since last century, we have witnessed a increasing frequency in the number and severity of droughts in particular territories such as South Africa *SA* [**?**]. The consequences here can be very dramatic. Due to the fact that area such as *SA* suffer from low economic power, a drought may drive towards dramatic consequences It can affect either agriculture, for a period that lasts up to 6 months, either hydrological up to 24 months [**?**].

There have been several studies showing how to reduce the impact of drought on the agriculture field by growing restraint to higher temperature and computational model to predict droughts [**?**].

However, in this paper, we wish to focus our attention on the drought consequences on wildlife population evolution. For instance, different studies have shown as animals mortality grows during the highest temperature period within the driest areas [**?**]. We will expand this idea to more recent data about waterbirds [**?**]. This data set will provide the number of individual for different species in each year within the last 45 years. By doing so, we are able to track the evolution of the species during time. However, the data set will not provide the number of new death, new born and the relative causes. It will be our job to build a model which allows us to investigate the relation other factors. In particular the data we are going to analyse regard the global temperature and rain fall amount [**?**] from 1901 up to 2020 which would give a foundation to build on the model of interest.

Having access to this information, our goal is to build a model which relates the climate change to the number of individual for each species and predict how, following the current trend, these would evolve over time and the potential consequences these could drive towards.

In section 2 we give a brief introduction to the methodology used within this research project. In section 3 we present the data with an exploratory data analysis. In section 4 we model the climate data and provide prediction on a possible future scenario. In section 5 we model the species count in relation to the climate data and provide prediction about the future number of individuals. In section 6 we discuss the finding of the project.

2 Technical Background

2.1 Generalized Lienar Model

In order to achieve the goal, we are going to employ different methodology which would allow us to model the data of interest. In particular, first of all we will related the temperature data to the rain fall amount. In order to do so, we will employ the Generalized Linear Model [**?**] which theory would allow to establish a linear relationship between the quantity of interest (rain) and a set of covariates (temperature) plush a stochastic term which models the uncertainty about the random variable realisation

$$g(\mathbf{Y}) = \mathbf{X}\beta + \epsilon$$

where we define a distribution $\mathcal{D}$ of interest on the stochastic term ϵ

$$\epsilon \sim \mathcal{D}(\theta)$$

and a link function $g(.)$ which is distribution dependent.

The model will be validate by a set of diagnostic on the residuals $\mathbf{Y} - \hat{\mathbf{Y}}$ which provide meaningful insight on the goodness of fitness.

2.2 Generalized Additive Model

We will further extend the *GLM* [**?**] theory to allow for non-linear relationship between the predictors $\mathbf{X}$ and the response $\mathbf{Y}$ exploiting the *Generalized Additive Model* theory [**?**] while maintining the model linear in the parameters

$$g(\mathbf{Y}) = f(\mathbf{X}) + \epsilon$$

where we define a distribution $\mathcal{D}$ of interest on the stochastic term ϵ

$$\epsilon \sim \mathcal{D}(\theta)$$

and a link function $g(.)$ which is distribution dependent. Moreover, we will define a function $f(.)$ which usually is a non linear function such as a *spline* [**?**].

2.3 Weighted Least Square

During the research project, we perform different model estimation and to remedy to inadequate diagnostics, such as those violating the homoskedasticity assumption, we will employ *Weighted Least Square* estimation theory [**?**] which allows us to iteratively estimate a weight matrix $\mathbf{W}$ to be employed in the model estimation as a measure of the importance for each unit $\mathbf{x}_i$ $i = \{1, \dots, N\}$ where N is the total number of observations.

2.4 Local Regression

2.5 Mixed Effect models

2.6 Ordinary Differential Equation

To track the evolution of a phenomenon over time, it is natural to think about *Oridnary Differential Equation (ODE)* theory [**?**] which allows us to keep track of the change of a quantity in continuous time

$$\frac{d}{dt}x = f(x)$$

given initial condition $f(x, t = 0) = x_0$

2.7 Levenberg-Marquardt algorithm

Very often, the function defined within an ODE is characterized by a set of parameters β. In empirical studies,we often have access to data realization $\{(y_i, \mathbf{x}_i,)\}_{i=1}^n$ we want to find the parameters β so that $y = f(\mathbf{x}, \beta)$ best fits the data of interest. The *Levenberg-Marquardt* [**?**] provide a useful and efficient solution to the problem by making smart combination of *Gauss-Newton* and *Gradien Descend* theory.

2.8 Sensitivity Analysis

Mathematical modeling is often subject to very strong assumption which may be difficult to evaluate. In order to provide a robust framework, it is necessary to validate the model under perturbations on the findings. This is why, we will exploit *Local Sensitivity Analysis* theory [**?**] to evaluate the robustness of the estimated models.

3 Exploratory Data Analysis

3.1 Rainfall amount

Within this exploratory data analysis context, we give a short introduction to the data by illustrating some characteristics. In the first place, we investigate the rainfall data set. We have access to the information between January 1901 up to December 2020. For simplicity, we decide to compute an annual average and inspect the resulting distribution.

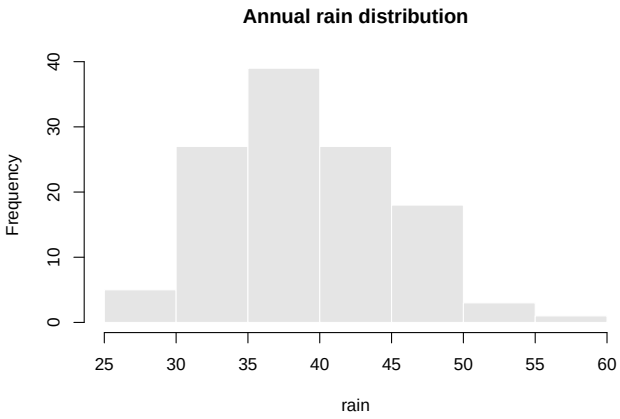


Figure 1: Yearly rainfall distribution

From figure 1 we can observe that the data distribution is somewhat bell-shaped. However, we can easily notice that the distribution is right skewed. Clearly, the phenomenon is left bounded by 0 $x \in \mathbb{R}_+$ and we can rarely observe very extreme event such as very heavy precipitation. We are able to quantify the amount of skewness by computing the data sample moments ratio known as the *Skewness* index which result to be 0.42, agreeing with the histogram depicted.

Afterwards, we investigate the temperature data. As for the rain data set, we have availability of the information from 1901 up to 2020. In figure **??** we depict the annual temperature distribution.

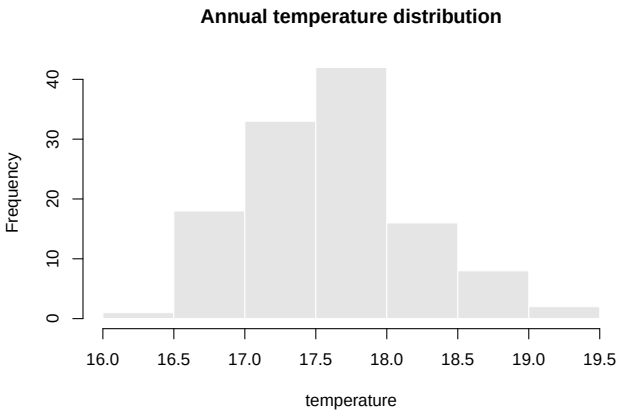


Figure 2: Annual temperature distribution

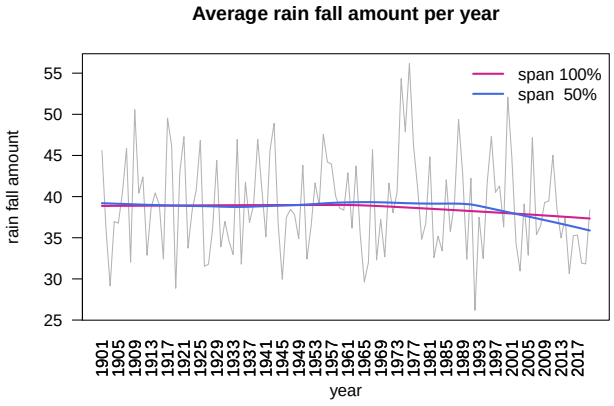


Figure 3: Annual rainfall

We can clearly observe a fairly strong right skeweness indicating more extreme events such as sweltering days than very cold days. The phenomenon now takes values in the real domain $x \in \mathbb{R}$ and present a skewness index of 0.43.

In order to give track the evolution of the phenomenon over time and to give an idea of the overall trend we make use of non parametric model, meaning that we do not assume anything but a relation between $\mathbf{x}$ and $\mathbf{x}$ governed by a function $f(.)$. The method employed is the locally weighted regression discussed in section 2.

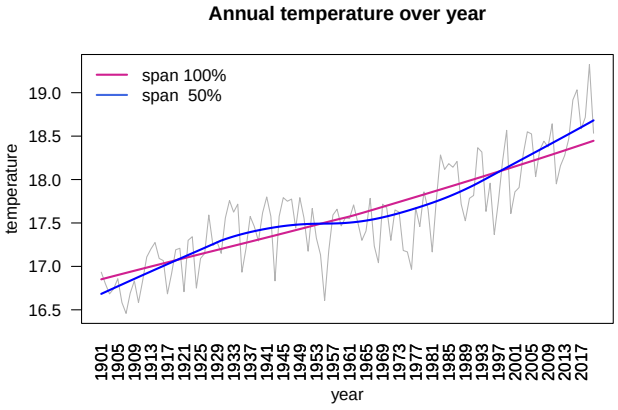
Therefore, to track the evolution of the phenomenon we fit the model for different values of the span. In particular:

- 100% of the neighbor data to take into account the global context (violet)

- 50% of the neighbor data to allow for a more local structure (blue)

By investigating the results in **??** we can notice how the global fit to the rain data shows a constant trend up to the 1950 which evolves in a slightly linear decreasing trend onward. On the other hand, the local fit shows a constant trend up to the 1980s which evolves in a steeper decrease for the later year. Nevertheless, both the models suggest an evolution of the amount of rainfall which changes over time.

Afterwards, we fit a the same two models for the temperature data. Here the relation with time is much clearer. For instance, we can observe how the global fit shows a linearly increasing trend over time, with constant velocity. On the other hand, the local fit shows an increasing trend with different intensities in different period allowing for non linear relationship. Nevertheless, the overall trend does not change dramatically. Annual temperature has clearly been increasing over the last century.



In order to give a global idea of the data we are investigating we would like to present the pairs plot showing each pair combination plus the estimated correlation matrix $\mathbf{R}$ in table **??**. As mentioned earlier, in figure **??** we can notice the strong relationship between temperature and year $\hat{\rho}_{t,y} = 0.82$ and a weaker relationship between rain and year $\hat{\rho}_{r,y} =$ -0.06 . However, we can see a clear relationship between year and rain $\hat{\rho}_{r,t} =$ -0.33 which may be worth further investigation.

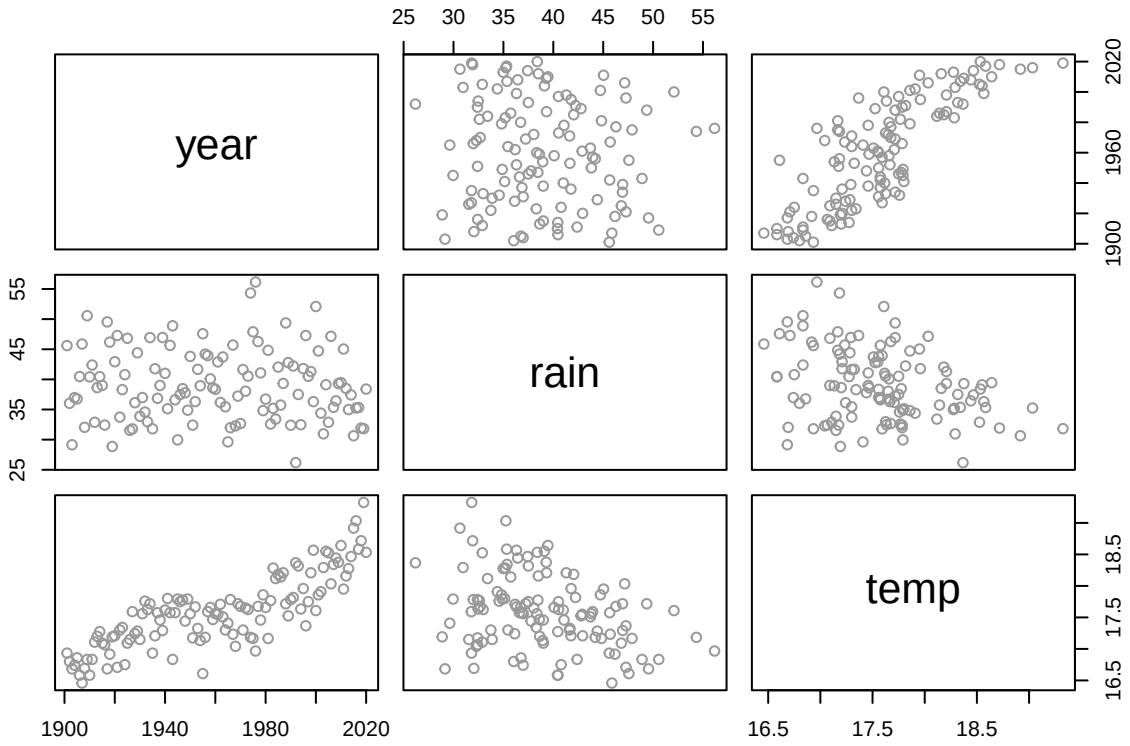


Figure 4: Year, Rain, Temperature scatterplot

4 Climate Modeling

We are now going to propose a very simple physical model to describe the temperature behavior. Ascertained that temperature heavily evolves over time, we present a very simple ordinary different equation given the temperature τ and time t

$$\frac{d}{dt}\,\tau = \gamma\,\tau$$

which posit a linear evolution of the phenomenon over time. Moreover, we add initial condition equal to the average temperature values of the first 5 year. The reason behind this choice is that, within this project, we do not have access to older data and during the exploratory data analysis in section refsec:eda we have seen how the phenomenon is highly variable. Therefore, we choose the mean as initial condition

$$f(\tau, t = 0) = \tau_0 = 16.8$$

Then we need to tune the parameter γ and we are going to do so exploiting the Levenberg-Marquardt algorithm discussed in section 2.

The algorithm applied to the data set of interest returns a point estimate of $\hat{\gamma}$ =7.6528×10^{-4} and an estimated standard error of $\hat{SE}(\hat{\gamma})$ =2.43×10^{-5}, resulting in the 95% confidence interval

$$P(\gamma \in [0.00071764, 0.00081291]) = 0.95 \qquad (1)$$

Overall, we can claim that, according to the model of interest, the global temperature increases over time by a factor within the region in (1). We graphically present the estimated model in figure 5. From the plot we can easily notice how the mean temperature increases over time. It also very noticeable how the model may underfit the data. It could be worth trying to fit a more complex model including higher order terms to address a certain amount of non-linearity. However, while these models may outperform the current in the data representation, they may be very unstable when employed in extrapolation task. The final goal here is to produce prediction about the future and a non-linear model may lead to very erratic conclusion. Therefore, we decide that this simple model fairly represents the phenomenon of interest.

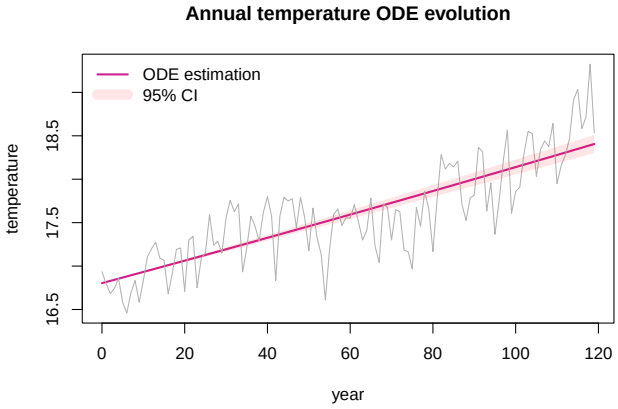


Figure 5: Annual temperature modeled by Ordinary Differential Equation

We have now available a model which may lead us to future prediction. In particular, we are interested in what would happen in a 50 years time following the current trend. According to the estimated model and depicted in figure 6, the global temperature may continue to rise up to reaching 19 Celcius average degree in 50 years time, with a reasonably tight confidence interval.

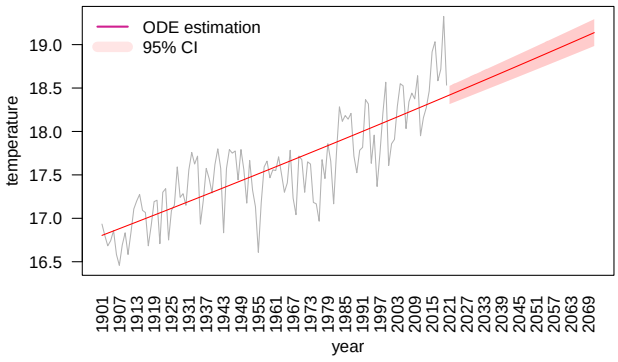


Figure 6: 50 out of sample years temperature prediction

A fundamental requisite for a mathematical model is the robustness to its input. Here, the data is empirically observed and the parameter γ is estimated from it. The question we want to address is whether we can rely on this parameter and whether the model output would change if the parameters in input changed. In order to tackle this question, a common and reliable choice is local sensitivity analysis discussed in section 2. Given the model

$$f(\tau) = \gamma\ \tau$$

the method allow us to derive the sensitivity equation as

$$\begin{aligned} \frac{d}{dt}s &= \frac{d}{d\gamma}f\ s + \frac{d}{d\gamma}f \\ &= \gamma s + \tau \end{aligned} \tag{2}$$

Therefore, we will need to deal with the system of ordinary differential equation

$$\begin{cases} \frac{d}{dt}\tau = \gamma\tau \\ \frac{d}{dt}s = \gamma s + \tau \end{cases} \tag{3}$$

As done previously, we numerically integrate the system of *ODE* and inspect the result. In particular, we want to investigate what would happen with different values for the input parameter γ. In this setting, we try to evaluate the model output for values which are 5 time the estimated standard error away from the point estimate $\gamma \pm 5 \times \hat{SE}(\hat{\gamma})$. In figure 7 we can observe the system output. In the first place, plotting the temperature values τ against time t in log-scale, we do not observe any appreciable difference for such an extreme value of the parameter considered. In the second plot, we plot the sensitivity value s against time t and, as before, we cannot notice any significant difference. This analysis clearly state how the model of interest is robust to perturbations of the input.

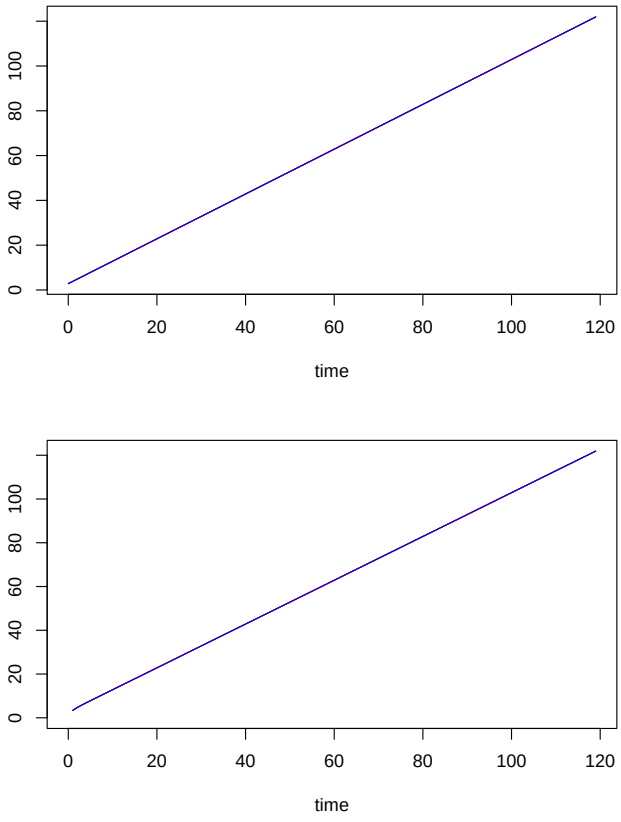


Figure 7: Evolution of temperature in time with input parameter $\pm$ 5 $\hat{SE}(\hat{\gamma})$; b) evolution of γ parameter in time $\pm$ 5 $\hat{SE}(\hat{\gamma})$

Once ascertained the robustness of the temperature model, we are ready to investigate between the latter and the rainfall amount. In order to do so, we estimate a linear regression model where we add a non-linear component derived by the spline theory which allows us to write

$$\mathbf{rain} = \beta_0 + f(\tau) + \epsilon$$

The first trial is not satisfactory as the model present some deficiency. In particular, from the left plot in figure **??** we can observe the presence of heteroskedasticity, meaning that the variance is not constant over the range of values for the value in input. In order to address the problem, we perform an iterative weighted least square estimation gaining a 29% reduction in the estimated BIC values, as shown in figure**??**.

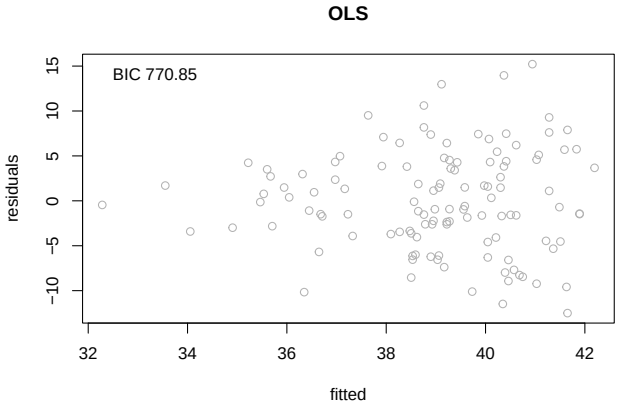


Figure 8: Residual vs Fitted plot for Ordinary Least Square model estimattion

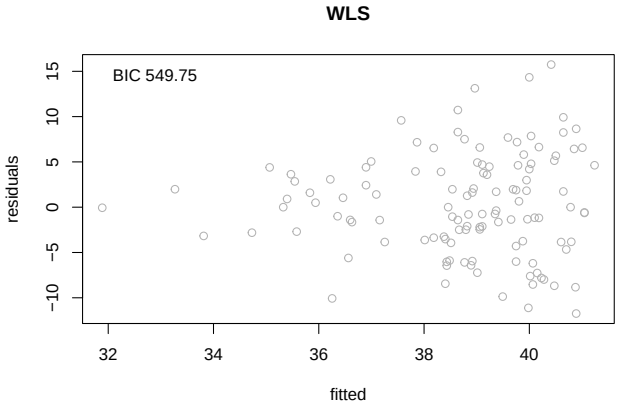


Figure 9: Residual vs Fitted plot for Weighted Least Square model estimattion

The model parameters results are summarised in table **??**. Unsurprisingly, from those we can observe how both the estimated spline coefficients have negative sign with very tiny estimated standard error. Therefore, we can appreciate how, up to 8 decimal precision, the returned P-value is 0, meaning that we reject the null hypothesis H_0 : $beta_i = 0$ for $i = \{0, 1, 2\}$. This means that, according to the estimated model, the relation between temperature and rainfall amount is negative. With higher temperature we can expect to observe lower amount of rainfall and vice versa.

	Estimate	Std. Error	t value	Pr($>|t|$)
$\hat{\beta}_0$	41.23907927	0.00008380	492090.20447958	0.00000000
$\hat{\beta}_1$	-7.90378231	0.00015488	-51030.55292079	0.00000000
$\hat{\beta}_2$	-8.53033020	0.00001818	-469180.40392581	0.00000000

Table 2: Weighted Least Square model estimatio for temperature and rainfall amount

We now need to evaluate the robustness of the model with respect to the parameters in input. As we did previously we will try two now configuration where we shift the parameters estimated by 5 times their standard deviation $\hat{\beta}_i \pm \hat{SE}(\hat{\beta}_i)$. The results are depicted in figure **??**. Here can notice how the different scenario have different intensities but the overall trend is very similar leading to likewise trend in the long period.

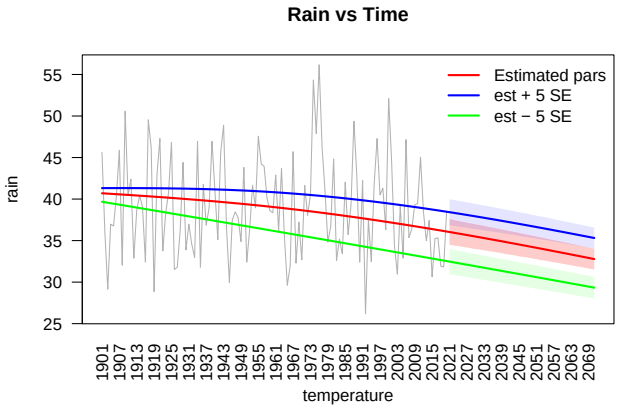


Figure 10: Evolution of rainfall w.r.t. temperature with input parameter $\pm$ 5 $\hat{SE}(\hat{\gamma})$

********** write conclusion here **********

5 Species count Modelling

In the previous section we have presented the data and results on the climate data including temperature and rainfall information from 1901 up to 2021. In this section, we are going to present the waterbirds species count data over time and relation its evolution to the covariate explored earlier. In particular, the data set of interest provide the specie individual count per year from 1975 to 2021. Within this data set we observe monthly variation which, for simplicity and consistency with the climate data, will be averaged over the years. For different species, we have different number of observation and, as we observe up to 37 species, we decide to keep only those whose numerousness might provide robust result, from $n_i = 30$ onward, as per figure **??**.

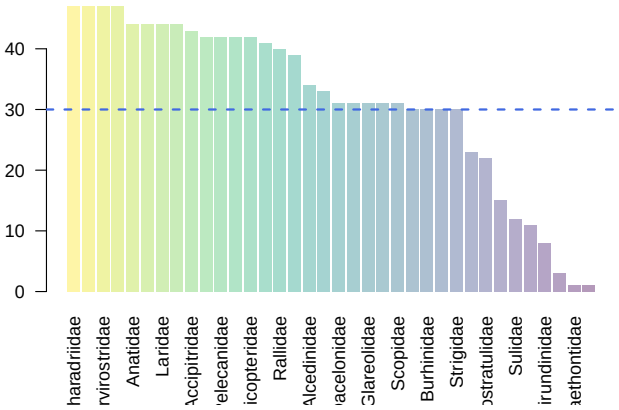


Figure 11: Years observation per species

Once selected the species of interest, we are interested in the overall trend. Beside a descriptive analysis, we provide a non-parametric fit by a local weighted regression to give an idea of the overall trend. As the values variance is appreciable large, we provide either mean (red line) and median (blue line) value for robust result. Despite being on different scale, both the results shows similar overall trend. The individual count has been decreasing over time. This behaviour is shown in figure **??**.

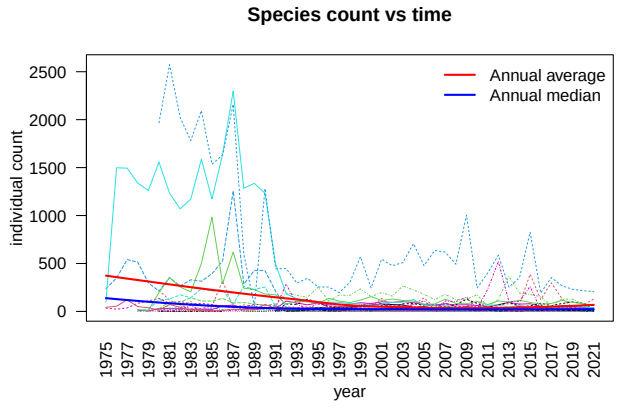


Figure 12: Yearly individual count per species plus mean and median observation

The next step is to relate the individual count data to the climate information. In the first place, we inspect the response distribution, as shown in figure **??**. As the data takes value on a discrete and positive set, it is reasonable to assume a *Poisson* distribution.

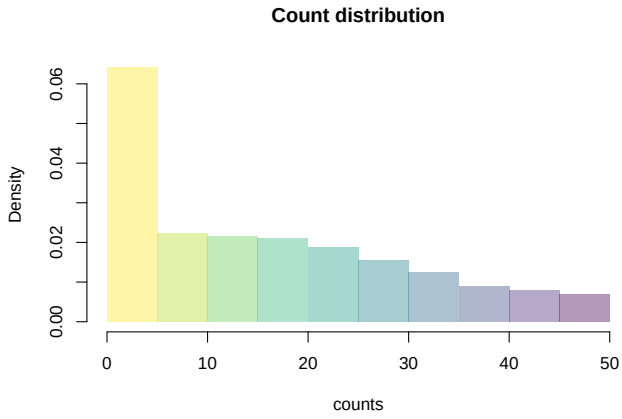


Figure 13: Distribution for animal counts

We now want to estimate a model describing the evolution of the population w.r.t. the climate input feature. In order to do so, we need to employ mixed-effect models since its theory allows us to include data from different class, namely their species, and thus taking account for the within correlation. The model to be estimated is of the form

$$\mathbf{count}_i = \alpha_i \, \mathbf{Z}_i + f(\mathbf{X}) + \epsilon$$

where $\epsilon \sim Poi(\lambda)$ and the term α indicates a random intercept whose contribution depends on the $\mathbf{Z} \in \mathbb{N}$ matrix which indicates the species belonging

$$z_{ij} = \begin{cases} 1 \text{ if observation i is in species j} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

thus varying the intercept of the model according to the species of interest. The function $f()$ employed in the estimation is, as usual, a set of spline which have empirically proven to be optimal for their high flexibility and low computational cost. The related degree has been chosen by making usage of the BIC measure which aims to maximize the likelihood while taking into account a penalization for the number of parameters estimated. The optimal value has empirically shown to be $K = 4$.

As we are including different predictors in the model, in order to get stable results we apply a pre-processing step which standardize the variables to have mean zero and variance one

$$Z = \frac{X - \mu}{\sigma}$$

From the model results presented in table **??**, we can observe how the coefficients related to the *year* marginal term $\hat{\beta}_{i,year}$ are all negative. Moreover, the coefficients related to its interaction terms $\hat{\beta}_{i,year,j}$ are negative as well. On the other hand, we can easily identify positive coefficients for the marginal terms $\hat{\beta}_{temp}$ and $\hat{\beta}_{rain}$. However, their magnitude is relatively small compared to the terms effect they interact with $\hat{\beta}_{i,year,temp}$ and $\hat{\beta}_{i,year,rain}$, therefore, we expect an overall negative effect.

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	4.21	0.30	13.89	0.00
$\hat{\beta}_{1,year}$	-2.36	0.03	-71.72	0.00
$\hat{\beta}_{2,year}$	-0.40	0.04	-10.50	0.00
$\hat{\beta}_{3,year}$	-1.17	0.08	-14.77	0.00
$\hat{\beta}_{4,year}$	-2.71	0.05	-56.05	0.00
$\hat{\beta}_{temp}$	0.75	0.04	20.37	0.00
$\hat{\beta}_{rain}$	0.68	0.02	30.32	0.00
$\hat{\beta}_{1,year,temp}$	-0.08	0.04	-2.17	0.03
$\hat{\beta}_{2,year,temp}$	-1.06	0.04	-24.22	0.00
$\hat{\beta}_{3,year,temp}$	-1.85	0.09	-20.58	0.00
$\hat{\beta}_{4,year,temp}$	0.08	0.04	1.95	0.05
$\hat{\beta}_{1,year,rain}$	-0.30	0.03	-11.41	0.00
$\hat{\beta}_{2,year,rain}$	-0.85	0.04	-23.99	0.00
$\hat{\beta}_{3,year,rain}$	-2.00	0.06	-33.35	0.00
$\hat{\beta}_{4,year,rain}$	-0.59	0.06	-10.40	0.00

Table 3: Mixed effects model of the individual count on the temperature and rainfall data results

For the model of interest, we have assumed the error distributed as a *Poisson* distribution, therefore, in the fitted vs residuals inspection, we run into a different behavior. In particular, in figure **??** we can observe how the majority of the fitted values of very small number close to zero with very few extreme values. Nevertheless, we can still see some structure in the data. For instance, below a 500 values for the fitted values, we can clearly see that the majority of the residuals are below the zero, while fewer are above zero. This diagnostic indicates that the model may be further improved by the inclusion of other predictors which, for this particular study, haven't been included.

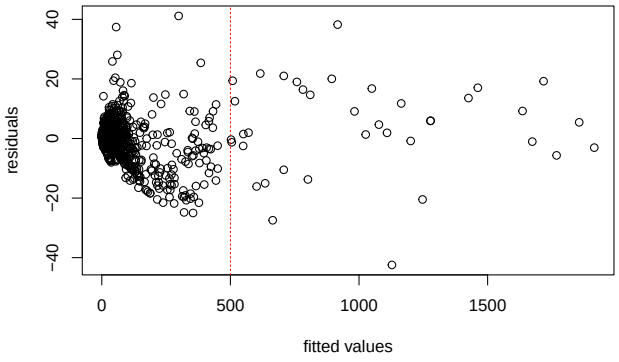


Figure 14: Distribution for animal counts

To make easier, the model comprehension, we will now subset the data to one of the species with the highest numerousness and focus the model analysis exclusively on it. In particular, we extract the data about the *Charadriidae* specie. This specie in particular is at high risk 15 a), reporting very low number in the last years, as per figure 15 b).

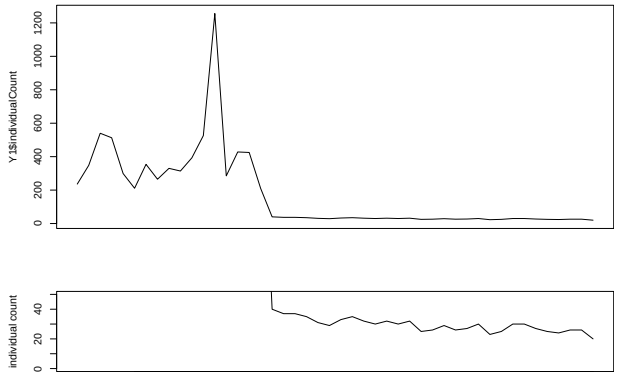


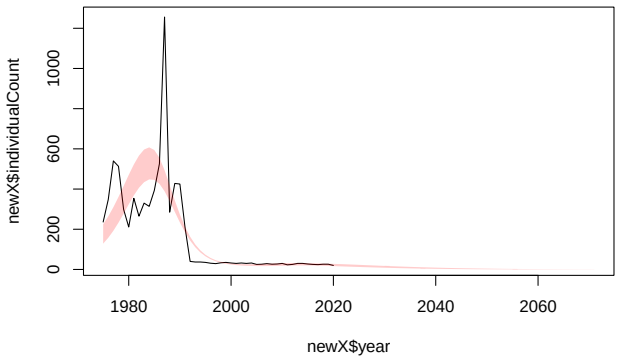
Figure 15: Charadriidae species count a) with focus on last years b)

To establish the relationship between the predictors and the response, we are going to make use of the *GLM* theory assuming, as we did previously, a *Poisson distribution*

$$\textbf{count} = f(\mathbf{X}) + \epsilon \quad \text{where } \epsilon \sim Poi(\lambda)$$

where $f()$ is a set if splines. The model results are shown in table **??**

	Estimate	Std. Error	z value	Pr(>|z|)
$\hat{\beta}_0$	5.31	0.09	60.03	0.00
$\hat{\beta}_{1,year}$	-2.48	0.10	-23.80	0.00
$\hat{\beta}_{2,year}$	-2.71	0.19	-14.36	0.00
$\hat{\beta}_{3,year}$	-0.40	0.24	-1.63	0.10
$\hat{\beta}_{4,year}$	-3.31	0.26	-12.56	0.00
$\hat{\beta}_{temp}$	0.22	0.09	2.48	0.01
$\hat{\beta}_{rain}$	0.42	0.06	7.40	0.00
$\hat{\beta}_{1,year,temp}$	-0.65	0.13	-5.13	0.00
$\hat{\beta}_{2,year,temp}$	0.47	0.24	1.99	0.05
$\hat{\beta}_{3,year,temp}$	-0.04	0.26	-0.14	0.89
$\hat{\beta}_{4,year,temp}$	-0.56	0.23	-2.50	0.01
$\hat{\beta}_{1,year,rain}$	-0.67	0.09	-7.21	0.00
$\hat{\beta}_{2,year,rain}$	-0.01	0.19	-0.07	0.95
$\hat{\beta}_{3,year,rain}$	-0.66	0.22	-2.96	0.00
$\hat{\beta}_{4,year,rain}$	-0.67	0.30	-2.22	0.03



6 Conclusion

References

[1] M. R. . Y. Richard, "Intensity and spatial extension of drought in South Africa at different time scales," *African Journals Online*, vol. 29, no. 4, 2003.

[2] M. Grayson, "Agriculture and drought," *Nature Outlook*, vol. 501, no. 7468, 2013.

[3] J. C. H. et. al, "Mortality of wildlife in Nairobi National Park, during the drought of 1973-1974," *African Journal of Ecology*, vol. 15, no. 1-18, 1977.

[4] G. G. B. I. Facility, "Free and open access to biodiversity data." www.gbif.org/occurrence/.

[5] W. B. Group, "Climate change knowledge portal." https://climateknowledgeportal.worldbank.org/download-data.

[6] P. McCullagh and J. Nelder, *Generalized Linear Models*. International series of monographs on physics, Chapman and Hall, 1989.

[7] R. T. Trevor Hastie, *Generalized Additive Models*. 1990.

[8] M. W. M. Aertsa, G. Claeskensb, "Some theory for penalized spline generalized additive models," *Journal of Statistical Planning and Inference*, vol. 103, no. 455-470, 2002.

[9] P. M. Hooper, "Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models," *Journal of the American Statistical Association*, vol. 103, no. 88:421, 2012.

[10] G. Gbur, *Mathematical methods for optical physics and engineering*. 2011.

[11] M. J.J., *The Levenberg-Marquardt algorithm: Implementation and theory*. Lecture Notes in Mathematics, Springer, 1978.

[12] J. Morio, "Global and local sensitivity analysis methods for a physical system," *European Journal of Physics*, vol. 32, no. 6, 2011.