Introduction to Statistical Data Science (STAT0032)

University College London

Department of Statistics

Group 2 Report

Authors

16098023

18027527

19074642

20029348

21134873

December 13, 2021

1 Introduction

Forest fires are common in Portugal; there have been 7,114 forest fires between January and October 2021, resulting in 26,833 hectares (ha) of burnt area. [1]. Although this was the lowest number of fires and burnt areas since 2011, there is large inter-annual variation, and the European Environmental Agency notes that there will be an increased risk due to climate change. [2]. Portugal is particularly at risk due to its high forest area; In 2010, Portugal had 801kha of natural forest, extending over 24% of its land area. In 2020, 16.8kha of this was lost. [3].

Although forest fires are a part of a natural cycle, they can have dire consequences. They may cause damage to land property and risk to human life. In 2017, wildfires killed 64 people in Portugal and destroyed 30,000 ha. [4]. Early detection and prediction are keys to mitigating the risks associated with forest fires. [3]. The ability to accurately predict forest fires will help manage scarce firefighting resources; in 2021, Portugal had access to 12,058 operatives, 2,795 teams, 2,656 vehicles, and 60 aerial means on the ground. [1]

This report provides a statistical analysis of the forest fire data from the northeast of Portugal, as taken from Cortez et al. [3]. Improving our understanding of such fires allows us to use advanced modelling techniques to predict future forest fires. This may allow the Portuguese government to manage firefighting and prevention resources accordingly.

2 Discussion of the Problem

In order to model, predict and assess the risk of forest fires, we first need to identify the distribution of the area destroyed by forest fires. We focused our analysis on August and September as these months had the highest number of fires, and the weather conditions during these months increased the difficulty of controlling fires (measured using the "Daily Severity Rating"). [5, p. 65]. Knowing the distribution of the forest area burnt during these months will help the Portuguese government estimate the probability of forest fire occurrences and the specific ranges of burned areas.

Additionally, it is essential to know if the area destroyed by forest fires in these two months follow the same distribution. Identifying this will help local authorities deduce the optimal number of firefighting resources that need to be allocated to each area in a given month. These resources include emergency services, educational campaigns, support and reconstruction of affected communities, preventive measures, and fire management decisions.

We used hypothesis testing to address our questions of interest. We first conducted an exploratory data analysis to identify critical aspects relevant to our analysis. Then based on our observations, we conducted goodness-of-fit tests to evaluate if the area destroyed by forest fires in the summer (August and September) followed a log-normal distribution and employed two-sample tests to assess if the area burned by forest fires during these two months followed the same distribution. Finally, we concluded with practical recommendations for the Portuguese government and stated the limitations of our statistical analysis.

3 Presentation of Dataset

The dataset used consisted of forest fires in the Montesinho natural park situated in Trás-os-Montes (province in the northeastern region of Portugal) between January 2000 and December 2003 [3]. Although the dataset contains 13 variables, we only considered the variables month and area for our analysis. We discarded fires with an area equal to or smaller than 0.01 ha as these are usually inconsequential.

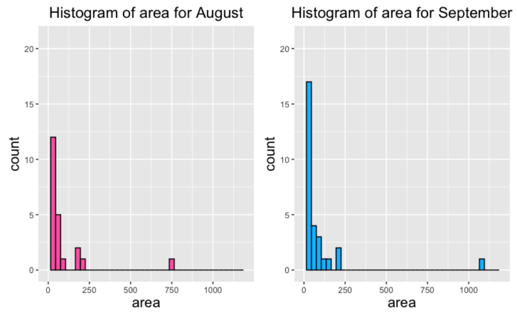

Figure 1: Histograms for area for August and September, 2000-2003

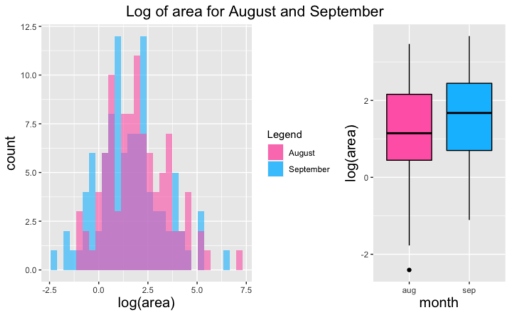

Figure 2: log of area for August and September, 2000-2003

The histograms in Figure 1 reveal that the area destroyed by forest fires during August and September are heavily positively skewed, with the mean values (23.21 for Aug and 31.82 for Sep) exceeding the median (4.40 for Aug and 6.58 for Sep).

The histogram in Figure 2 illustrates that the logarithm (with respect to base e) of the area destroyed by forest fires appears to follow normal distribution for both August and September. The boxplots for August and September vs log(area) in Figure 2 suggests that both months may have a similar distribution. Although the interquartile ranges for both boxplots were very similar, the quartiles for August (0.57, 1.48, 2.40) were lower than the quartiles for September (0.78, 1.88, 3.00).

4 Goodness-of-fit tests

We used the Anderson-Darling test and the Shapiro-Wilk test to test the following hypotheses:

- H_0: The forest area burnt (> 0.01 ha) per month follows a log normal distribution
- H_1: The forest area burnt (> 0.01 ha) per month follows a different distribution

We will use a 1% significance level (vs standard 5%) as type 2 errors (failing to reject H_0 when it is false) are less consequential in this analysis. As mentioned in section 3, even if the area destroyed by forest fires does not perfectly follow a log-normal distribution, it may still be good enough for modelling purposes. If the p-value obtained from the tests is less than 0.01 (1%), there is sufficient evidence to reject H_0. Otherwise, we will proceed under the assumption that H_0 is true.

4.1 Anderson-Darling test

For this test, the data is standardised using $Y_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}$ where X_i is forest area burnt & $\hat{\mu}$ and $\hat{\sigma}$ are estimates of the mean and standard deviation respectively. The test statistic used in this test is: $A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}((2i-1)ln\Phi(Y_i) + ln(1-\Phi(Y_{n+1-i})))$ where Φ is the standard normal CDF (Cumulative Distribution function)$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Y_i} e^{-t^2/2}dt$, and n is the number of data points. We used estimates of the mean and variance so that the p values are calculated from the modified statistic $a = a_\infty(1 + \frac{b_0}{n} + \frac{b_1}{n^2})$ where a_∞, b_0 and b_1 are given in Table 4.9 in [6] as per R documentation. [7].

4.2 Shapiro-Wilk test

This test uses the quantiles to assess if a random sample of the data is normally distributed.

The test statistic is calculated as follows: $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ where $x_{(i)}$ is the ith order statistic (the ith smallest number in the sample) and $\bar{x} = \sum_{i=1}^{n} x_i$ is the sample mean. The coefficients a_i are given by $(a_1, ..., a_n) = \frac{m^T V^{-1}}{C}$ where m is a vector consisting of the expected values of the order statistics of independent and identically distributed (iid) random variables sampled from the standard normal distribution, V is the covariance matrix of these normal order statistics and C is the vector norm $C = ||V^{-1}m|| = (m^T V^{-1}m)^{\frac{1}{2}}$. [8] The test statistic W, does not follow a known distribution and the critical region for the test is determined using Monte carlo simulations. [8]

4.3 Tests and results and comparison

The results of both tests are shown in tables 1 and 2. The p-values for both tests for each month are greater than 0.01, so there is insufficient evidence to reject the null hypothesis that the forest area burnt (> 0.01 ha) during August and September follow a log-normal distribution. We will proceed under the assumption of a log-normal distribution.

Table 1: Anderson-Darling test results.

Month	Test Statistic: A	P-value
August	0.407	0.3425
September	0.317	0.5343

Table 2: Shapiro-Wilk test results.

Month	Test Statistic: W	P-value
August	0.98797	0.5132
September	0.98524	0.3511

Both tests assume that all samples are drawn from the same distribution and are independent (iid). i.e. the area of a fire does not depend on previous fires. The Shapiro-Wilk test has been shown to have the highest power of any normality test in most cases. However, Shapiro-Wilk tests only for normality and it requires a computer due to its complexity.[9]. In comparison, Anderson-Darling is marginally less powerful but can perform better in some cases when an alternative distribution is specified. The Anderson-Darling test does not work for categorical data and its statistic is based on squares. [10]

5 Two sample tests

We used two non-parametric tests; the two-sample Kolmogorov-Smirnov (KS), and the two-sample Anderson-Darling (AD) test to test the following hypotheses:

- H_0: The fire area (> 0.01 ha) in August and September follow the same distribution.
- H_1: The fire area (> 0.01 ha) in August and September have different distributions.

We will use a significance level of 5%, as type 1 and type 2 error will be equally consequential (both result in allocating incorrect number of firefighters in August or September).

5.1 Two-sample Kolmogorov-Smirnov test

The idea behind this test is to compare the empirical CDF of each sample and derive the maximum distance (D) measure between the two samples as a test statistic with the formula shown below [11]: $D = max|F_a(x) - F_b(x)|$ where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,x]}(X_i)$ is the empirical CDF of the sample n at time point x. This test statistic follows a Kolmogorov distribution. One key benefit of this test is that it is a non-parametric test. This means that it does not depend on the data following a known distribution. [12]. As the empirical CDF is obtained by summing the indication function of the observations and normalizing, this will be sensitive to the shape and tails of the distribution.

5.2 Two-sample Anderson-Darling test

This test has the advantage of not having to assume a specific distribution. It is a non-parametric test in which we look at the ranking between two samples to generate a test statistic. The test statistic is $AD2 = \frac{1}{mn}\sum_{i=1}^{k-1}\frac{(kc_i - mi)^2}{i(k-i)}$, where m & n represents the sample size of the two ordered sample, k is the total number of combined samples, and c_i is the number of combined elements that are smaller than the i^{th} sample [13].

5.3 Test results and comparison

Table 3: Two-sample test results.

Test used	Test Statistic	P-value
Two-sample KS test	0.176	0.083
Two-sample AD	1.24	0.099

The results are shown in Table 3. At the 5% level of significance, we have insufficient evidence to reject the null hypothesis that the area destroyed by fires for the two months follow the same distribution. As we have strong evidence of normality from section 4, we considered running a two-sample t-test and an F-test for equality of variance to compare the mean and variance that characterizes the distribution between the two months. However, as these tests test only one parameter at a time, they may not be as appropriate as the non-parametric tests that we have used.

Both tests used assume that all samples are independent and come from a common unspecified continuous distribution (iid), but the continuous assumption for the AD test can be ruled out for discrete cases [14] [15, p. 6]. The two tests compare the CDF between the two samples to come up with a test statistic; nonetheless, the AD test is more powerful and requires less data to obtain sufficient statistical power than the KS test [16]. Therefore, the two-sample AD test seems to be the most appropriate test for our objective.

6 Conclusion

6.1 Benefits and advice

Our analysis showed there was insufficient evidence to reject the hypothesis that the area destroyed by fires during both August and September followed a log-normal distribution at the 1% significance level. There was also insufficient evidence to reject the hypothesis that the fire area for both August and September follow the same distribution, at the 5% level. Based on this, we suggest proceeding under these assumptions. Subsequent modelling can assume that the fire area follows a log-normal distribution. As stated in [3], early detection and prediction is the key to mitigating the risks associated with forest fires. Given that there was insufficient evidence to suggest August and September follow different distributions, we recommend allocating similar firefighting resources for both months.

6.2 Limitations

There are several limitations to our analysis. Firstly, all tests performed assume that the forest fires are independent (i.e. one fire is unrelated to previous fires). However, this may not be the case if previous fires deplete the available fuel for future fires by burning through a significant forest area. Further, we may have failed to reject the null hypotheses purely because we have insufficient data to do so. Moreover, the power of the test depends on the significance level selected. We have selected 1% and 5% as the levels of significance, based on the impact of type 1 and 2 errors. However, this is still a subjective choice, and we may have rejected the null hypotheses if we used less conservative tests (higher levels of significance).

Our report analysed the data for August and September between the years 2000 - 2003 (a subset of the data available). This data is 15 years old, so the conclusions drawn may no longer be relevant. To provide more accurate advice, we suggest analysing each month of a more up-to-date dataset to take into account seasonal variations and the distribution of fire areas for all months. Moreover, we have not considered other variables such as temperature, dryness, wind speed, etc., that may affect the area destroyed by fires. The above modelling improvements may allow the Portuguese government to allocate resources observing a higher degree of freedom.

References

[1] *Portugal: Critical wildfire period ends with lowest number of rural fires in last decade.* 2021. URL: https://aerialfiremag.com/2021/10/01/portugal-critical-wildfire-period-ends-with-lowest-number-of-rural-fires-in-last-decade/.

[2] *Forest fires in Europe.* 2021. URL: https://www.eea.europa.eu/ims/forest-fires-in-europe.

[3] Paulo Cortez and Aníbal Morais. "A Data Mining Approach to Predict Forest Fires using Meteorological Data". In: 2007.

[4] *Portugal forest fires under control after more than 60 deaths.* 2017. URL: https://www.theguardian.com/world/2017/jun/22/portugal-forest-fires-under-control.

[5] San-Miguel-Ayanz J et al. "Forest Fires in Europe, Middle East and North Africa 2019". In: KJ-NA-30402-EN-N (online),KJ-NA-30402-EN-C (print) (2020). ISSN: 1831-9424 (online),1018-5593 (print). DOI: 10.2760/468688(online),10.2760/893(print). URL: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC122115/jrc122115-annual_report_2019_final_topdf_1.pdf.

[6] R.B. and Stephens M.A. D'Agostino. *Goodness-of-Fit Techniques.* Marcel Dekker, 1986. ISBN: 0-8247-7487-6.

[7] *ad.test function - RDocumentation.* 2021. URL: https://www.rdocumentation.org/packages/nortest/versions/1.0-4/topics/ad.test.

[8] S. S. Shapiro and M. B. Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52 (3–4 Dec. 1965), pp. 591–611. ISSN: 0006-3444. DOI: 10.1093/biomet/52.3-4.591.

[9] Henry C. Thode. *Testing for normality / Henry C. Thode, Jr.* eng. Statistics, textbooks and monographs ; v. 164. New York: Marcel Dekker, 2002. ISBN: 0824796136.

[10] Edith Seier. "Normality Tests: Power Comparison". In: *International Encyclopedia of Statistical Science* (2011), pp. 1000–1003. DOI: 10.1007/978-3-642-04898-2_421. URL: https://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_421.

[11] W.J. Conover. *Practical Nonparametric Statistics.* Wiley, 1971, pp. 309–314. ISBN: 9780471168515.

[12] NIST/SEMATECH. *e-Handbook of Statistical Methods. Kolmogorov-Smirnov Goodness-of-Fit Test.* Tech. rep. National Institute of Standards and Technology, 2021. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm.

[13] A. N. Pettitt. "A Two-Sample Anderson–Darling Rank Statistic". In: *Biometrika* 63.1 (1976), pp. 161–168. ISSN: 00063444. URL: http://www.jstor.org/stable/2335097.

[14] F.-W. Scholz and Michael A. Stephens. "K-Sample Anderson-Darling Tests of Fit, for Continuous and Discrete Cases". In: 2008. URL: http://l.academicdirect.org/Horticulture/GAs/Refs/Scholz&Stephens_1986.pdf.

[15] Angie Zhu [aut] Fritz Scholz [aut cre]. *kSamples: K-Sample Rank Tests and their Combinations.* R package version 1.2-9. 2019. URL: https://cran.r-project.org/web/packages/kSamples/kSamples.pdf.

[16] Sonja Engmann and Denis Cousineau. “Comparing distributions: the two-sample Anderson–Darling test as an alternative to the Kolmogorov–Smirnov test”. In: *Journal of Applied Quantitative Methods* 6 (Sept. 2011), pp. 1–17. URL: `http://www.jaqm.ro/issues/volume-6,issue-3/pdfs/1_engmann_cousineau.pdf`.

7 Additional page

Student no. 16098023: I worked on the Two sample tests comparing forest fire data points between August and September, as well as coming up with a conclusion for the course of actions of our client. I am fully aware of the content of the ”Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science, and to the best of my knowledge, this report complies with those rules.

Student no. 18027527: I worked on the presentation of the dataset section and the second part of goodness of fit hypothesis tests. I feel we all contributed equally to this report. I am fully aware of the content of the “Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science, and to the best of my knowledge, this report complies with those rules.

Student no. 19074642: I contributed to sections 2, 3, and 5. I am fully aware of the content of the “Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science, and to the best of my knowledge, this report complies with those rules.

Student no. 20029348: I worked on the Two sample tests of our report, mainly the code part. I feel we all contributed equally to this report. I am fully aware of the content of the ”Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science, and to the best of my knowledge, this report complies with those rules.

Student no. 21134873: I worked on the introduction and first part of goodness of fit hypothesis tests. I feel we all contributed equally to this report. I am fully aware of the content of the “Plagiarism and Collusion” section in the Taught Postgraduate Student Handbook for the Department of Statistical Science, and to the best of my knowledge, this report complies with those rules.