

**Working With Data: MTHM501**

**Final Assessment**

**Declaration of AI Assistance**: I have used OpenAI’s ChatGPT tool in creating this report.

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

1. I have used GenAI tools for developing ideas.
2. I have used GenAI tools to help me understand key theories and concepts.
3. I have used GenAI tools to check and correct my code.
4. I have used GenAI tools to suggest a plan or structure for my assessment.
5. I have used GenAI tools to proofread and correct grammar or spelling errors.
6. I have used GenAI tools to give me feedback on a draft.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.

To what extent can spatial and temporal patterns accurately classify whether a fire is a war-related fire in Ukraine?

**Introduction:**

The conflict between Ukraine and Russia escalated sharply on February 24, 2022, when Russia launched a full-scale invasion, intensifying hostilities that began with the 2014 annexation of Crimea and the rise of separatist movements in Eastern Ukraine. Since then, over 30,000 civilians have been killed or injured, with 3.7 million displaced internally and 6.5 million fleeing the country, marking one of Europe’s most severe humanitarian crises. Beyond the human toll, widespread attacks on military and civilian areas have caused massive infrastructure damage. Hundreds of fires, many linked to the conflict, occur weekly. Accurate classification of these incidents is essential for emergency response, assessing the conflict's impact, and aiding long-term reconstruction.

**Research Objectives**

The primary objective of this analysis is to determine if spatial and temporal patterns can be used to classify fires in Ukraine as war-related. I will explore three sub questions:

1. Are certain regions or seasons more related with war-related fires?
2. Can factors like population density and sustained excess be reliable predictors for classifying fires as war-related?
3. Which patterns provide the most accurate and reliable predictions to assess a fire’s classification?

**Hypotheses**

**Null Hypothesis (H0):** *Spatial and temporal factors do not influence the classification of fires as war-related.*

**Alternative Hypothesis (H1):** *Spatial and temporal factors are significant predictors of war-related fires.*

**Report Structure**

This report includes a section on hypothesis testing, followed by exploratory data analysis, detailed hierarchical model results, a discussion of key findings, implications and limitations, and a conclusion.

**Data:**

The primary dataset used for this analysis, ‘ukraine\_fires.csv,’ was sourced from Kaggle and created by The Economist’s Ukraine War-fire model. It includes over 60,000 observations with 30 parameters, capturing geospatial ('latitude', 'longitude'), temporal ('date', 'AQC time'), and contextual variables ('pop\_density', 'sustain\_excess'). Additionally, a GeoJSON file with regional mapping of Ukraine was merged, allowing for refined geospatial analysis.

**Data Wrangling and Preprocessing**

*Can spatial, temporal, and contextual variables accurately predict whether a fire event is war-related in Ukraine?* With this broader question in mind, I begun wrangling the data.

**Data Cleaning**: I began with tidyverse and dplyr packages for initial cleaning, using head(), str(), and summary() to assess the dataset. Using Amelia's missmap(), I identified that the 'city' parameter was entirely null, leading me to remove both ‘city’ and ‘year’ parameters, along with other irrelevant fields. I checked for duplicates and refined the dataset by removing eleven additional unnecessary parameters and renaming the remaining for clarity.

**Data Formatting**: To reformat dates, I applied lubridate, creating a ‘season’ variable by extracting the month and grouping the 'time\_of\_day' variable into morning, afternoon, evening, or night. I also converted the key categorical variable, ‘war\_fires,’ into a factor to facilitate analysis.

**Geospatial Context**: Using the sf package, I prepared to merge the dataset with Ukraine's regional map. After validating the spatial variables and converting the dataset into an sf object, I merged the regions column and set region and season as factors to enable random effects modelling.

**Exploratory Data Analysis**

Table 1.

A screenshot of a graph

Description automatically generated**Summary Table of War-Fires by Region and Area**

The "Summary of War-related Fires by Region" table reveals clear regional patterns in war-related fire activity. Sumy and Chernihiv show the highest concentrations (East and Northern), with war fires comprising 75.3% and 70.4% of total fires, indicating intense conflict impact. In contrast, Kiev and Zaporizhia have lower proportions of war-related fires at 24.8% and 57%, suggesting varied conflict exposure. Western regions like Chernivtsi and Zakarpattia show minimal or no war-related fires, likely due to their distance from conflict zones.

These findings indicating spatial factors significantly affect the likelihood of a fire being a war-related fire. Data handling used dplyr for summarization and gt for table styling.

Figure 1.

**Spatial Visualisation of War-related Fires**

**A map of ukraine with red dots

Description automatically generated**Figure 1(gif), generated using gganimate and ggplot2, depicts the distribution of war-related fires across Ukraine from the start of the war in February 2022 to October 2024. Each red mark represents a war-related fire event, while grey denotes non-war fires, allowing for clear visual differentiation. I used ggspatial to add a distance scale and ggshadow to mark past events. These packages combined allow enable advanced visualisation, and deeper insight.

As shown in Figure 1, we can see a high concentration of war-related fires around 100km from Russia’s border, especially in Eastern Ukraine, with much lower activity as you move west. Temporally, marked surges in mid-2023 and throughout 2024 suggest recurring conflict intensity patterns. These spatial and temporal trends support rejecting the null hypothesis, confirming their significance in predicting war-related fires. The next analysis will cover the visible seasonal patterns. (Find the animated gif on my GitHub – linked in appendix)

A graph of different types of fires

Description automatically generated with medium confidence**Seasonal Trend Plot with War-related fires, Showing Peaks in Mid-2023 and Late-2024.**

Figure 2.

Figure 2 reveals that war-related fires peak between July and October annually, with noticeable increases each successive year. Reaching 2000 daily fires 2023, and 5000 2024. There is a consistent influx of war-related fires at the beginning of spring (April), suggesting higher likelihoods of conflict-related fire incidents during late Spring, Summer and Autum.. Another temporal pattern we see, the total amount of war-fires event is increasing year on year, indicating that as time passes, the probability of a fire being a war-related fire is higher.

The seasonal trends plot would support you to reject the null hypothesis as it illustrates how temporal factors, specifically seasonal shifts, significantly correlate with the frequency of war-related fires, indicating that certain seasons exhibit heightened conflict activity.

Figure 2 displays war-related fires per day over time, with distinctive colours representing each seasons. I adjusted the war\_fire\_per\_day variable, using the summarise and mean function, then grouping it by date and season. Next, I extracted the yearly data from date with lubridate. I utilised ggplot2 with grid, gridExtra, and dplyr for structuring. Putting the plots together provide us with greater insight into temporal patterns.

**Hypothesis Testing**

**Hierarchical Modelling**

Hierarchical modelling will be used, as it allows us to account for region-specific and seasonal differences in fire activity, capturing the nested nature of spatial and temporal data. Since my outcome variable is binary, I decided to use Generalised Linear Mixed Models (GLMM) and specifically using the logistic regression link, R’s lme4: :glmer, methods not covered in the course. As my outcome variable is binary, this improves the models fit, robustness and insights. The model includes *population* *density* and *sustained* *excess* *fire* as my fixed effects, with regional(25 levels), seasonal(4 levels), and coordinate-based clustering(106 levels) as random effects. I believe these this model, gives us a good approximation whilst staying relatively simple.

**Model Setup:**

I used `str()` to confirm variables were in the correct format to optimise the model. I converted random effect parameters to factors, and log-transformed, then standardised `population\_density` to improve interpretability, model convergence, and reduce multicollinearity.

**Model Specification**

1. **Random Intercept Mode (RIM)**: *glmer(war\_fire ~ population\_density + sustained\_excess\_fires + (1 | region) + (1 | season) + (1 | id\_big), data = TestData1, family = binomial)*

Assumes varying intercepts for regions and seasons.

1. **Random Slopes Model (RSM)**: *glmer(war\_fire ~ population\_density + sustained\_excess\_fires + (1 + population\_density | region) + (1 + sustained\_excess\_fires | season) + (1 | id\_big), data = TestData1, family = binomial)*

Allows varying intercepts and slopes, accounting for variations in population density and sustained fire activity across regions and seasons. This model was expected to provide better insights given the anticipated regional and seasonal fluctuations in fire occurrences.

**Results**

**Model Comparison and Analysis**

Table 2.

**A screenshot of a computer

Description automatically generated**

Table 2 shows that the RSM has a slightly lower AIC and BIC, indicating the model is a better fit. The Sustained Fire Activity Estimate remains highly significant across both models (p < 2e-16). Population Density is only significant in the RIM (p = 0.0012), suggesting that its predictive effect may vary when accounting for random slopes. We will cover the random effect variances later in the analysis.

Overall, the RSM provides a nuanced approach, supporting the alternative hypothesis that spatial and temporal factors significantly predict war-related fires. This table was formatted using knitr and kableExtra, and the modelling was conducted with lme4 in R.

**Model Diagnostics**

**A graph of a graph

Description automatically generated**

Figure 3.

The Binned Residuals Plots examines residual behaviour across different predicted probability ranges. This method is more suitable than a Q-Q plot or normality histogram, as my data is binary. A sample of 10,000 observations were used to streamline processing. In the RIM, the residuals exhibit a U-shaped pattern, especially at probability extremes, indicating systematic bias with underprediction or overprediction. In contrast, the RSM shows a flatter residual trend around zero in the midrange, suggesting a more accurate fit with fewer systematic deviations, supporting it as the better model for capturing variability across regions.

The Intraclass Correlation Coefficient (ICC) further reinforces this preference: the RIM has a lower ICC (0.113), while the RSM, with an ICC of 0.206, demonstrates slightly better reliability in accounting for regional differences in war-related fires. Figure 3 was created using ggplot2, organized with patchwork, and prepared with dplyr.

**Model Insights**

**A screenshot of a graph

Description automatically generated**

Figure 4.

The Random Effect Variance Plot demonstrates substantial regional and seasonal variability in the effects of population density and sustained excess fires. High-density regions such as Zaporizhia exhibit greater variance, suggesting that areas with larger populations may be more strategically targeted or impacted by conflict. Seasonally, the RSM highlights variability in sustained fires activity, indicating that environmental conditions or shifts in conflict intensity influence fire classification differently across seasons.

While fixed effects reveal a general association between high population density, sustained excess fires, and an increased likelihood of war-related fire classification. The random effects underscore that these influences vary across specific regions and seasons. This variation supports the use of the RSM, which captures these nuanced patterns more effectively than the Random Intercept Model. This choice of model, emphasising localised effects, gives us greater insight into predicting spatial and temporal patterns, and therefore, should provide a higher degree of accuracy when classifying fires. These observable variances also indicate the presence of spatial and temporal patters, defining their relationship to war-related fires further. Figure 4 was created using sjPlot, ggplot2, patchwork, and stringr in R. I log’ed the x-axis, so differences in variance are clearer.

**Model Suitability**

The RSM proves to be the more accurate model for testing the hypothesis on spatial and temporal patterns in classifying fires as war-related. By allowing predictor effects to vary across spatial and temporal groups, the RSM outperforms the RIM on key diagnostic assumptions, including linearity, homoscedasticity, and normality of residuals. This flexibility reduces bias in fitted values and provides a closer approximation of normal residual distribution, resulting in a more accurate representation of the data.

Additionally, the anova() function confirms the RSM’s superiority based on metrics such as AIC, BIC, log-likelihood, and deviance, with the Chi-squared test further supporting its improved fit. This enhanced performance makes the RSM a more robust model for capturing the spatial and temporal variability inherent in war-related fires, directly supporting the hypothesis that these factors are significant predictors.

**Hypothesis Testing**

**Bonferroni Correction**

To control for multiple comparisons, a Bonferroni correction adjusted the significance level to 0.0125 (initial α = 0.05; m = 4 comparisons). This is crucial, as it prevents type 1 error arising from multiple comparisons, thus keeping the models statistical integrity.

**Analysis**

The likelihood ratio test (LRT), using R’s lmtest package, comparing the null model (see code) with the RSM confirms that sustained fires activities is a highly significant predictor of war-related fire classification (Estimate = 16.4944, p < 2.2e-16). Population density does not show significance (Estimate = -0.1157, p = 0.266) at the adjusted level.

The model’s random effects reveal notable variance across regions and seasons, illustrated in the Random Effect Variance Plot, underscore the uneven impact of population density and sustained excess fires across regions and seasons. The binned residuals plots support rejecting the null hypothesis by demonstrating improved model fit with the RSM, which accounts for regional and seasonal variability. The flatter residuals trend suggests fewer systematic errors, confirming spatial and temporal factors as significant predictors of war-related fires. This result aligns with AIC, BIC, and Chi-squared test results from the anova(), reinforcing the model’s improved accuracy when these factors are considered.

**Conclusion**

The analysis supports rejecting the null hypothesis, confirming sustained excess fires as a significant factor in classifying war-related fires. Although population density was not significant, the random effects in the RSM highlight essential regional and seasonal variability. Therefore, the model confirm that spatial proximity to conflict zones and seasonal variations play critical roles in predicting and classifying war-related fires.

A chart of different colors

Description automatically generated

Figure 5.

Figure 5 (interactive) visualises the predicted probabilities of war-related fires from the RSM across regions and seasons, organised into four areas of Ukraine (East, North, South, West)(top to bottom). Two distinct patterns appear, aligning with the research question on spatial and temporal patterns.

Firstly, a spatial pattern is clear, with eastern and northern regions consistently showing higher predicted probabilities of war-related fires (around 0.65), compared to western regions, which average a probability of 0.1. This confirms that war-related fires are more concentrated in specific areas, those closer to Russia, supporting the hypothesis that spatial patterns influence war-related fire classifications.

Secondly, a clear temporal pattern shows that the probability of war-related fires steadily increases from winter to autumn across most regions, showing seasonal factors may heighten conflict intensity as the year progresses. Together, these patterns affirm that both spatial and temporal factors are significant predictors. Thus, supporting us to reject the null hypothesis.

Created using ggplot2’s geom\_tile(), plotly, lme4, stringr, tidyr, and dplyr. I obtained the predicted probabilities using the predict() function from the lme4 package. (Find the interactive plot on my GitHub – linked in appendix)

**Discussion:**

The analyses confirm that spatial and temporal patterns are significant predictors of war-related fires in Ukraine. The Random Slopes Model provided the best fit, effectively capturing the regional and seasonal variations that simpler models could not. Higher probabilities of war-related fires in Eastern and Northern Ukraine suggest that proximity to conflict zones substantially increases the likelihood of such events. Seasonally, an upward trend from winter to autumn indicates that fire occurrences may align with specific environmental conditions or intensified conflict periods.

**Implications**

These findings have practical implications for conflict and environmental management. By identifying high-risk regions and seasons, policymakers and emergency responders can better allocate resources for fire prevention and response. The results provide a foundation for more targeted monitoring and intervention in areas and times most affected by conflict-related fires.

**Limitations**

The binary sustained fire activity variable limits insight by not capturing the intensity or duration of fires, which could refine our understanding of fire dynamics. Additionally, cloud cover in satellite imagery likely obscured some fires, leading to underreporting and potential bias in the data.

**Future Research**

Future studies could improve on these findings by refining the sustained fire activity variable to capture more detailed fire event data and incorporating conflict-specific and socio-economic data for more nuanced modelling. Using ground-based observations to complement satellite data could also enhance the reliability of war-related fire analysis, reducing the limitations posed by cloud cover.

**Conclusion**

In answer to the research question—"To what extent can spatial and temporal patterns accurately classify whether a fire is a war-related fire in Ukraine?” The study finds that these patterns can indeed serve as accurate classifiers, strongly rejecting the null hypothesis. Spatial and temporal factors, as modelled through the Random Slopes approach, provide significant predictive power in distinguishing war-related fires, particularly when accounting for regional proximity to conflict zones and seasonal variations. This research underscores the value of spatial and temporal data in conflict-related environmental analysis and highlights avenues for refining predictive capabilities in future studies.