Using Historical Reports of UFO Sightings from NUFORC Dataset to Predict Future Sighting Trends

Shreyansh Misra, Lawrence Meng

2024-12-9

The NUFORC Databank is the largest independently collected set of unidentified flying object (UFO) sighting reports available on the internet. This investigation aimed to uncover trends and patterns in UFO sightings by addressing three primary questions:

1. **Where are UFOs most likely to be sighted?** Are they concentrated in specific countries or states?
2. **When are UFOs most likely to be sighted?** Are sightings tied to specific months or dates?
3. **What are the most common UFO descriptions?** What shapes are commonly reported?

By analyzing these aspects of UFO sighting reports, we can use the trends that we find in order to predict the location, timing, and descriptions of future UFO sightings. We aim to use a Time Series Forecasting model along with a Regression Model in order to carry out our predictive analysis.

1 Data Preperation

1.1 Overview

The dataset contains 80,332 records of UFO sightings, with variables detailing sighting locations, times, and descriptions. There are minimal missing variables as the NUFORC omits them from the dataset if there is not enough information. However, there are still some missing values that we adress on a case by case basis in the analysis section. It is also difficult to gauge outliers in this investigation since we are not assesing the accuracy of the reports, and are simply looking to find trends within them.

1.2 Variables

- **Datetime**: Date and Time of the sighting.
- **City/State/Country**: City, State, and Country of Sighting (eg. ashburn, nc, us).
- **Shape**: Reported shape of the UFO (levels discussed in Analyzing Descriptions).
- **Duration (seconds)**: Length of the sighting.
- **Latitude/Longitude**: Geographic coordinates of the sighting.
- **Comments**: Eyewitness accounts.

1.2.1 Variables of Interest

- **Datetime**
- **City/State/Country**
- **Shape**

Analyzing these variables will help us answer the questions: When do UFO sightings take place, where are they most frequent, and what are the most common descriptions?

2 Exploratory Analysis

2.1 Analyzing Location

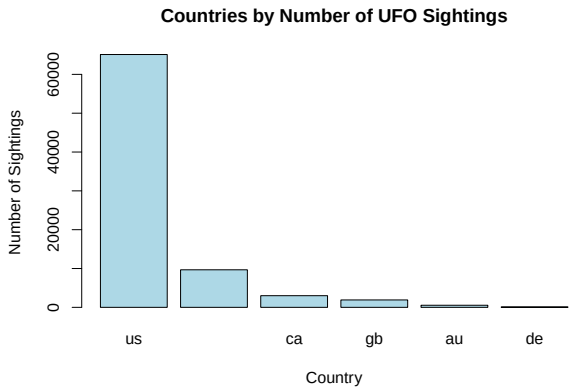

Figure 1: Countries by Number of UFO Sightings

The United States overwhelmingly has the highest number of UFO Sightings reported, followed by Canada, Great Britain, Australia, and Germany. There are also a large number of reports where the country is left blank. The NUFORC being based in the United States is likely why most of their reports are from the United States. Since the sightings from outside the US are too few to come to reasonable conclusions with, we will only use sightings based in the United States for this investigation.

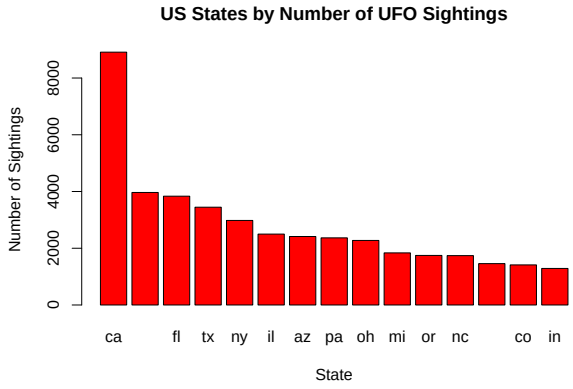

Figure 2: US States by Number of UFO Sightings

California had the most UFO sightings reported, followed by Washington, Flordia, Texas, and New York. It is significant to note that California, Texas, Florida, and New York are the four most populous states in that order. Furthermore, all of those states have a coastline. California in particular has a very diverse landscape with many remote areas and a large amount of "deserted" space without light pollution that can make it easier to spot usual objects in the night sky. Those states also have a very significant military presense and drone testing, fighter plane tests, and other military activities can often be mistaken for UFOs.

2.2 Analyzing Descriptions

There were 29 different "shapes" that the NUFORC Databank classified each UFO sighting into. People reported what UFO looked like and it was grouped into a category based on their descriptions. While they are labelled as "shapes" in the dataset, a more accurate way to define the data column would be "descriptions" of UFOs. This is because some data points include "fireball", "light", "flash", and "flare", which are good descriptions of what the UFO would have looked like but are not explicitly shapes.

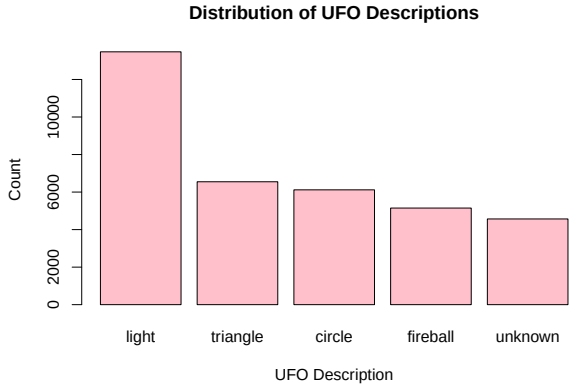

Figure 3: Distribution of UFO Descriptions, warning=FALSE

The top five "descriptions" of UFOs were light, triangle, circle, fireball, and unkown respectively. There is not much to analyze when it comes to UFO descriptions since there is a lot of variation in what people see, but it is significant to note that accounts of "light" and "fireballs" could be mistaken for shooting starts, comets, or airplane lights.

2.3 Analyzing Time

The dataset includes historical reports dating back to 1910. For the context of this investigation, we will be eliminating reports collected before 1990. This means that the oldest report was taken from "1990-01-03" and the most recent from "2014-05-01".

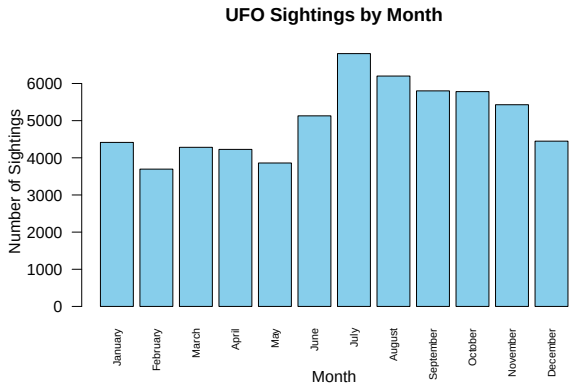

Figure 4: Distribution of UFO Sightings by Month

July, August, September, and October have the highest number of sightings.

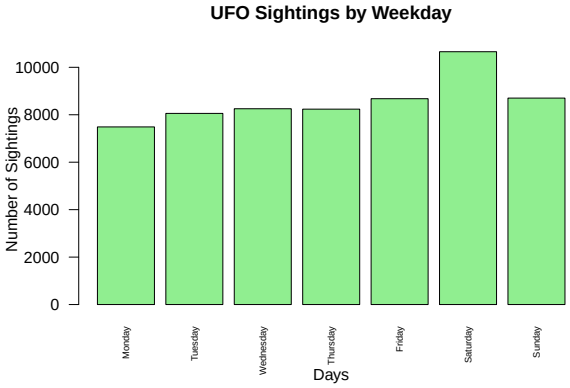

Figure 5: Distribution of UFO Sightings by Weekday

The most UFOs are spotted on Saturdays, followed by Fridays and Sundays.

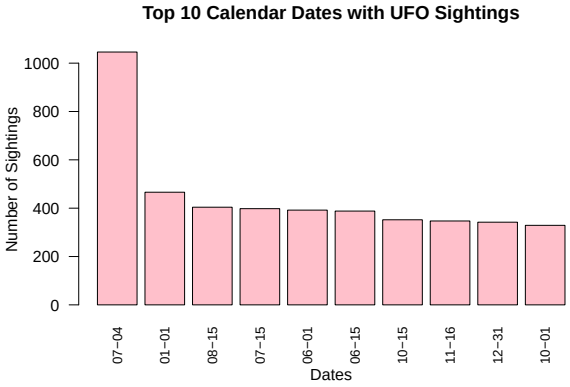

Figure 6: Top 10 Calendar Dates with UFO Sightings

The dates with the most UFO sightings across the 24 year period observed are the 4th of July, 1st of January, and the 15th of Auugust. The first two of these dates are days of festivities in the United States, namely the 4th of July (Independence Day) and New Years Day. These two days are also associated with high levels of alcohol consumption, which could explain the stark increase in the number of sightings on those days. Furthermore, it makes sense that the 4th of July has any more sightings than on New Years Day since people tend to spend the 4th of July outdoors.

This also puts the other observations into context. July, August, September, and October have the highest number of sightings since people are more likely to be outdoors during the evenings on those days. Furthermore, people are more likely to be drinking during the weekends, especially on Saturdays and Fridays, which explains the increase in the number of sightings on those weekdays.

3 Predictive Analysis

3.1 Time Series

From the exploratory analysis, time emerged as the primary predictor of UFO sightings, with clear evidence of seasonal patterns. A time series model is therefore an appropriate tool to forecast future sightings using

the NUFORC dataset. The target variable is the monthly number of UFO sightings, aggregated by year and month. The dataset spans a range of dates, from "1990-01-03" to "2014-05-01", with a monthly frequency. The time series object created reflects these attributes, providing a structured basis for analysis.

3.2 Exploration

We first processed the data to aggregate sightings by year and month, converting the datetime column into a monthly time series. Below is a plot of the time series, which shows trends, seasonality, and occasional spikes in sightings.

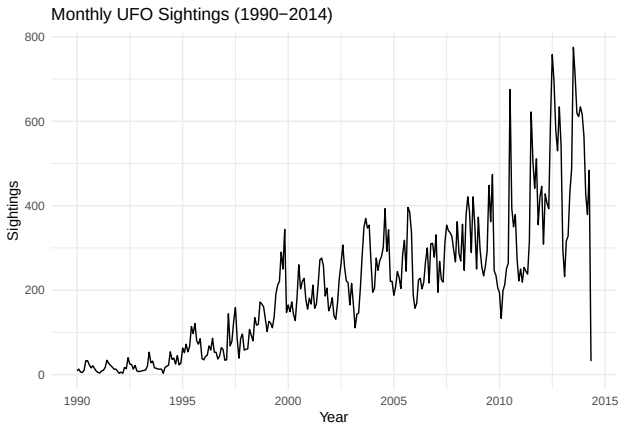

Figure 7: Monthly UFO Sightings (1990-2014)

This plot reveals: - Long-term trends: Increasing overall trend in UFO sightings. - Seasonality: Regular patterns of higher activity in certain periods or months.

3.3 Model Fitting

To forecast UFO sightings, we evaluated two models: - Benchmark Model: Mean forecast. - ARIMA Model: Captures trends, seasonality, and autocorrelations.

The dataset was divided into training and test sets: - Training Set: Covers all data except the final 12 months. - Test Set: The last 12 months, reserved for model evaluation.

3.3.1 Mean Forecast

The mean forecast assumes future sightings will equal the historical mean of past data. This provides a simple baseline for comparison.

3.3.2 ARIMA Model

The ARIMA model captures complex temporal patterns, automatically adjusting for trends and seasonality. Differencing was applied to ensure stationarity, as confirmed by the Augmented Dickey-Fuller (ADF) and KPSS tests.

- **Augmented Dickey-Fuller (ADF) Test:** -7.0054
- **KPSS Test:** 4.1877

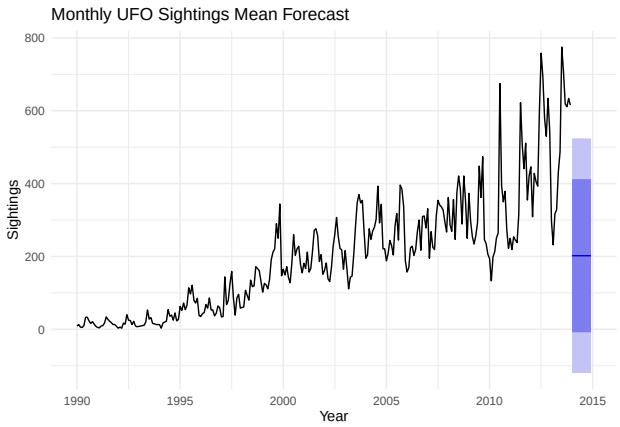

Figure 8: Monthly UFO Sightings Mean Forecast

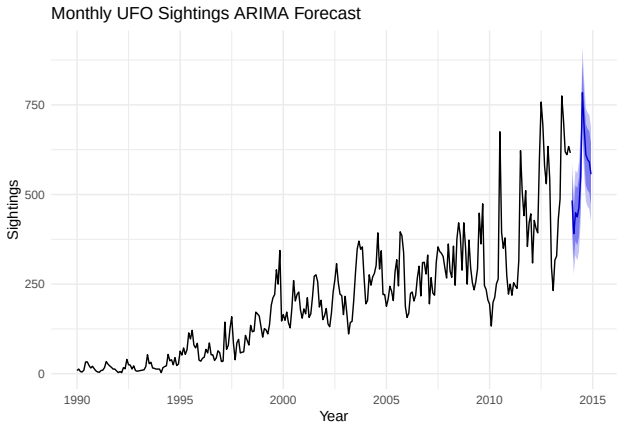

Figure 9: Monthly UFO Sightings ARIMA Forecast

3.4 Model Comparison

The models were compared using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), calculated on the test set.

Table 1: Performance Metrics for Mean Forecast and ARIMA

Model	MAE	RMSE
Mean Forecast	243.1875	253.5293
ARIMA	133.2323	200.9817

The ARIMA model outperformed the benchmark model in both MAE and RMSE, demonstrating its ability to account for trends and seasonality.

3.5 Best Model Fit and Forecast

The ARIMA model, selected for its superior performance, was refitted to the entire dataset. Predictions were generated for a 12-month forecast horizon.

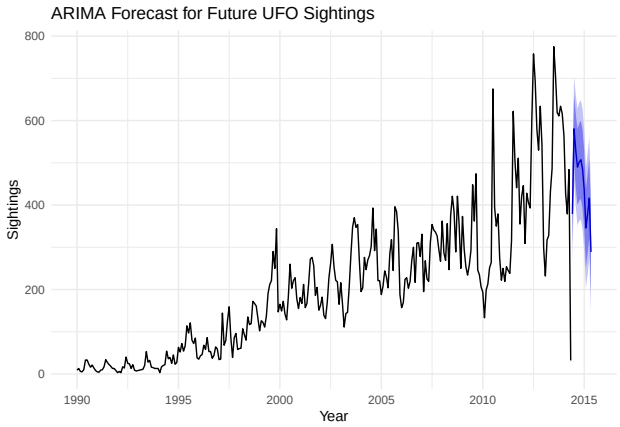

Figure 10: ARIMA Forecast for Future UFO Sightings

4 References

1. https://nuforc.org/databank/
2. https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/UFOsandUAPs/2d_af_1.pdf
3. https://attheu.utah.edu/facultystaff/the-west-is-best-to-spot-ufos/