Bericht der Gruppe „Ufo“

Eine Analyse der Abhängigkeit der Wetters auf vermeintliche Ufo-Sichtungen.

Lars Thomsen, Uzeyir Mammadov
Martin-Luther-Universität Halle-Wittenberg
[lars.thomsen][uzeyir.mammadov]@student.uni-halle.de

Abbildung 1: NUFORC.

Tabelle 1: Kennzahlen des Datensatzes.

Einträge	Anzahl
Gesamt	96 924
Davon einzigartige Orte	25 234
Orte mit Wetterdaten, gesamt	2 020
Davon mit Sonnenminuten pro Stunde	26
Davon mit Sonnenminuten pro Tag	116
Davon mit condition codes	1 886

ABSTRACT

KEYWORDS

Übung „Big Data Analytics“, Sommersemester 2021, Ufo

1 EINLEITUNG

Auf den ersten Blick klingen vermeintliche Ufo-Sichtungen immer nach an den Haaren herbeigezogenen Erfindungen und Lügengeschichten. Doch wie viel Wahrheit steckt in diesen Sichtungen, und lassen sich gegebenenfalls Muster in diesen erkennen?

Im Zuge unseres Projektes haben wir uns folgende Forschungsfrage gestellt: Existieren Abhängigkeiten zwischen vermeintlichen Ufo-Sichtungen, und welchen Einfluss nimmt das zum Zeitpunkt der Sichtung herrschende Wetter?

Der Bericht gliedert sich wie folgt: Kapitel 2 beschreibt die verwendeten Daten und Datensätze. In Kapitel 3 wird die Forschungsfrage und ihre Bearbeitung vorgestellt. Kapitel 4 diskutiert die Ergebnisse unserer Frage. Im Schlusskapitel geben wir unser Fazit und runden den Bericht ab.

2 DATEN

Die verwendeten Datensätze stammen aus einer Datenbank vom „National UFO Reporting Center“ (NUFORC). In diese Datenbank kann jede Person ihre vermeintliche Ufo-Sichtung – entweder online über ein Formular oder per Telefon – eintragen. Des Weiteren senden gewisse Messstationen (z.B. MADAR Nodes) auffällige Messdaten automatisch an die Datenbank. Um offensichtlichen Fakes entgegenzuwirken werden die eingereichten Daten vor der monatlichen Veröffentlichung von den Betreibern des Portals überprüft.

Jeder Eintrag des Datensatzes besteht aus dem Datum und der Uhrzeit der Sichtung, der Stadt sowie dem Kürzel des dazugehörigen Bundesstaates (ausschließlich für die USA), einer klassifizierten Beschreibung der Form des Ufos und gegebenenfalls einer kurzes Zusammenfassung und Beschreibung aus der Sicht des Einsenders.

Für diese Daten haben wir eine einfache Datenbank erstellt, um auch offline auf diese zugreifen zu können. Um die Daten einfacher verarbeiten zu können, bereiten wir diese während der Abfrage von unserer Datenbank auf und treffen eine Vorauswahl an brauchbaren Daten. Die Kriterien für die Vorauswahl wurden durch stichprobenartige Abfragen festgelegt. Dem entsprechend werden Einträge von vorhin beschriebenen externen Messstationen, Einträge mit fehlenden Inhalten oder „?“ als Inhalt nicht betrachtet. Der Datentyp aller Attribute ist der Einfachheit halber nur string. Zu einem späteren Zeitpunkt werden Datum und Uhrzeit in ein datetime-Format überführt. Mit dem Wissen aus den stichprobenartigen Abfragen legen wir und auf die zwei gängigsten Formate fest – 'm/d/y H:M' und 'm/d/y'. Andere in dem Datensatz vorkommenden Formate oder von den Einsendern eigenständige Angaben beachten wir nicht. Die aufbereiteten Daten werden letztendlich durch das Tripel `datetime`, `city` und `state` beschrieben.

Als Quelle für die Wetterdaten haben wir uns letztendlich für „Meteostat“ entschieden[2]. Auf diese Daten wird über die dazugehörige Python-Library zugegriffen. Essenziell für unser Projekt sind die Attribute `tsun` (Anzahl der Sonnenminuten) und `coco` (Klassifizierter Zustand des Himmels). Die Daten sind für unsere Vorhanben bereits eine ausreichende Qualität, sodass in diesem Fall keine weitere Aufbereitung und Säuberung der Daten nötig ist.

3 DIE ERSTEN VERSUCHE

Die erste Idee war es, die Wetterdaten für unsere Forschungsfrage von den „National Centers for Enviromental Information“ (NCEI) zu beziehen. Diese Datensätze beinhalten für unsere Analyse zwei wesentliche Attribute: `HourlySkyCondition` und `HourlyVisibility`. Diese beiden Werte beschreiben zum Ersten den klassifizierten Zustand des Himmels (z.B. Regenwolken, Schleierwolken, Nebel etc.) und zum Zweiten die prozentuale Bedecktheit des Himmels. Die Liste aller verfügbaren Stationen sind ebenfalls beim NCEI verfügbar(Link Stationen).

Um die Daten einer Ufo-Sichtung mit den dazugehörigen Wetterdaten zu verknüpfen, wird das jeweilige eindeutige `city, state`-Tupel der am nahestehendsten Wetterstation zugeordnet. Um die Ergebnisse nicht zu sehr zu verfälschen, wurde das zulässige Einzugsgebiet einer Wetterstation auf 20 Kilometer, mit der Station als Zentrum, begrenzt.

Diese Idee konnte leider nicht umgesetzt werden. Es standen zwar zwei Versionen der API vom NCEI zur verfügung, allerdings greifen diese auf unterschiedliche Stationen zurück. Dabei waren zwei Probleme die Hauptursache für das Aufgeben dieser ersten Idee: Zum Ersten waren die internen Bezeichnungen der Stationen bei Version 1 der API verschieden zu denen der uns zur verfügung stehenden Liste[4]. Eine einheitliche Liste aller Stationen, welche durch Version 1 der API bedient werden war nicht verfügbar. Zum Zweiten wurde von Version 2 der API eine verschiedene zur bereits verarbeiteten Liste an Stationen verwendet, welche nur unter Umständen einsehbar war[5]. Somit war es nicht möglich mit den verwendeten Datensätzen die gestellte Forschungsfrage zu bearbeiten.

4 FORSCHUNGSFRAGE

Nachdem wir uns dazu entschlossen haben, die APIs des NCEI nicht mehr zu verwenden, haben wir uns auf die Suche nach neuen APIs und Anwendungen gemacht, um an für uns passende Wetterdaten zu kommen. Der Service von Meteostat passte dabei perfekt in unsere Planung. Die Abfrage von Wetterdaten erfolgte durch die Python-Library von Meteostat und bekam als Input den Zeitraum der gewünschten Daten, sowie den Längen- und den Breitengrad des gewünschten Ortes.

Um die entprechenden geographischen Koordinaten für die Abfragen zu erhalten, wurde die API von „MapQuest“ verwendet. Als Parameter wurden das `city, state`-Tupel aus dem bereinigten Datensatz übergeben. Den Output bildeten ein Tupel aus den resultierenden Längen- und Breitengraden der gewünschten Orte. Um redundante Anfragen zu vermeiden, wurden zunächst einzigartige `city, state`-Paare aus dem Datensatz extrahiert. Somit reduzierte sich die Anzahl der zu bewältigenden Geocoding-Anfragen von 96 924 auf 25 234.

Mit den abgeschlossenen Geocoding-Anfragen waren alle benötigten Daten für die Anfragen an die Meteostat-Datenbank lokal verfügbar. Als Basis für die Abfragen galten die Daten aller Sichtungen „ `sightings.csv` “ mit dem Tripel `datetime, city, state`. Über das Tupel `city, state` wurden die entsprechenden Längen- und Breitengrade aus „ `cities_coords.csv` “ für die entsprechende Anfrage als Punkt gespeichert. Als Richtwert für den Zeitpunkt der abzufragenden Wetterdaten gilt das Attribut `datetime`. Da Meteostat für jede Anfrage einen Start- und Endzeitpunkt benötigt, wird als Endzeitpunkt eine Sekunde auf den Startzeitpunkt addiert.

Für die eigentliche Abfrage der Wetterdaten wurden zwei varianten benutzt. Zum Ersten sich stündlich aktualisierende Daten und zum Zweiten sich täglich aktualisierende Daten. Der Ablauf der Anfrage ist in beiden Fällen identisch, der einzige Unterschied ist, dass die Condition Codes für den Himmel nur in den stündlichen Daten vorhanden waren. Da die Stationen nicht identisch sind, lieferte nicht jede Station für jedes relevante Attribut die entsprechenden Messwerte. Es wurden dementsprechend nur die Ergebnisse der Anfragen gespeichert, wenn einer der drei relevanten Werte vorhanden war. Die Ergebnisse der Anfrage bestanden letztendlich aus der Liste `city, state, time_date, sun_hourly, sun_daily, condition_code`, welche im letzten Schritt visualisiert wurden.

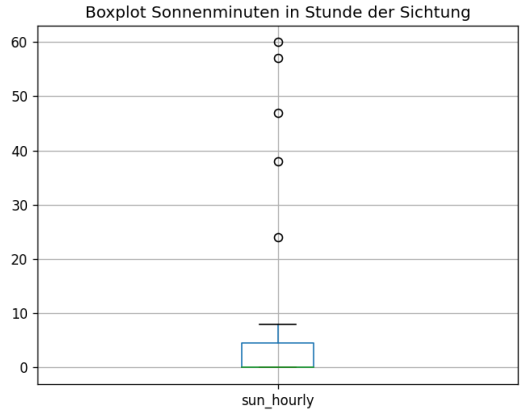

Abbildung 2: Sonnenminuten pro Stunde.

Die Visualisierungen erfolgten alle über die „matplotlib“-Library. Visualisiert wurden die drei Messwerte als Graph, die Condition Codes als Histogramm und Tortendiagramm sowie die stündlichen und täglichen Sonnenminuten als Boxplots.

5 EVALUATION

Die Bearbeitung der Forschungsfrage hat folgende Ergebnisse hervorgebracht: Aus ursprünglich 96 924 potenziell relevanten Ufo-Sichtungen im Datensatz konnten für 2 020 Sichtungen zu dem Zeitpunkt passende Wetterdaten gefunden werden, das entspricht ungefähr 2,1%. Mit welcher Art der drei Vergleichsdaten die Sichtungen analysiert wurden lässt sich in Tabelle 1 ablesen.

Die Vergleichsdaten mit dem am Abstand wenigsten Vorkommen bilden die Sonnenminuten pro Stunde mit lediglich 26 Einträgen. Wie in Abbildung 2 zu erkennen ist, befindet sich die Mehrheit davon im Bereich von 0 bis 5 Sonnenminuten pro Stunde. Vereinzelte Ausreißer reichen gegen 50 bis 60 Sonnenminuten.

Abbildung 3 beschreibt die Verteilung der Sonnenminuten pro Tag an jeder verfügbaren Ufo-Sichtung. Die mittleren 50% der Ergebnisse sind hierbei, im Gegensatz zu den stündlichen Sonnenminuten, breiter verteilt. Sie reichen von 200 bis 700 Sonnenminuten pro Tag mit wenigen Ausreißern, welche nur gegen weniger Minuten streben. Im Schnitt scheint die Sonne während eines Tages, an dem ein vermeintliches Ufo gesichtet wurde, um die 500 Minuten – also etwas mehr als 8 Stunden. In Anbetracht dessen, dass die durchschnittlichen Sonnenstunden in den USA im

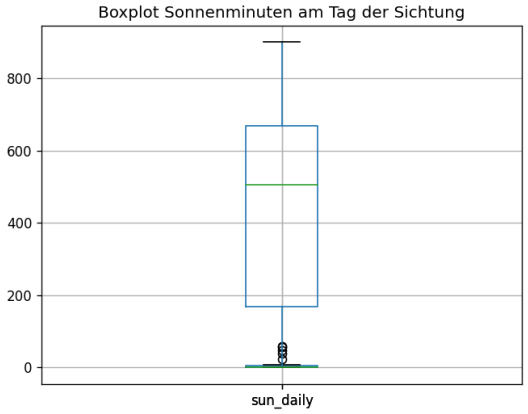

Abbildung 3: Sonnenminuten pro Tag.

Bereich zwischen 2 Stunden im Winter und bis zu 12 Stunden im Sommer reichen, kann man den Ufo-Sichtungen eine leichte Überdurchschnittlichkeit an Sonnenstunden während eines Sichtungstages zuordnen[6].

Das Attribut mit den am meisten verwendbaren Vergleichsdaten bilden die Condition Codes mit 1 886 Daten zur entsprechenden Sichtung. Auf ihnen liegt in dieser Evaluation das Hauptaugenmerk und die beiden Attribute der Sonnenminuten ergänzen diese. Aus Abbildung 4 geht hervor, dass sechs der 17 ermittelten Condition Codes überwiegen. Schönwetter „`Fair`“ dominiert die absolute Häufigkeit an Einheiten mit einer Anzahl von 603. Die genauen Werte können der Tabelle 2 entnommen werden. Im folgenden Abschnitt werden lediglich die sechs Attribute mit der größten Ausprägung (Codes 1, 2, 3, 4, 5, 7). Zusammengerechnet umfassen diese Condition Code-Gruppen 1 691 aller Ufo-Sichtungen – das entspricht 89,7%. Diese Codes können weiter in zwei Gruppen eingeteilt werden: Zum Ersten „gutes Wetter“ mit `Clear`, `Fair` und `Cloudy` und zum Weitern „schlechtes Wetter“ mit `Overcast`, `Fog` und `Light Rain`. Die Bezeichnungen können [3] entnommen werden. Damit die Gruppe „gutes Wetter“ auf 1 193 Einheiten und die Gruppe „schlechtes Wetter“ auf 498 Einheiten. Die restlichen nicht in Gruppen eingeteilten Codes umfassen vor allem weitere Stufen von Regen- und Schneeschauern sowie Gewitter. Generell kann man sagen, dass je höher der Condition Code ist, desto „schlechter“ ist das Wetter und umso schlechter kann man Objekte am Himmel erkennen. Die Codes sind also ordinal skaliert mit einem Median von $\tilde{x} = 3$. Ausgehend vom Median, welcher sich innerhalb der Gruppe „gutes Wetter“ befindet, gibt es überdurchschnittlich mehr Sichtungen mit gutem als mit schlechtem Wetter.

Für die Interpretation dieser Werte gibt es zwei unterschiedliche Herangehensweisen. Zum einen ist es ersichtlich, dass es bei gutem Wetter und freiem Himmel mehr Möglichkeiten gibt, Flugobjekte zu entdecken. Dies würde die Dominanz der Gruppe „gutes Wetter“ erklären. Es erweckt allerdings auch den Gedanken, wieso bei vermeintlich schlechterem Wetter, im Vergleich zu den nicht in Gruppen

Tabelle 2: Condition Codes.

Code	Weather Condition[3]	Anzahl
0	-	15
1	Clear	227
2	Fair	603
3	Cloudy	363
4	Overcast	93
5	Fog	186
7	Light Rain	219
8	Rain	51
9	Heavy Rain	9
12	Sleet	1
14	Light Snowfall	42
15	Snowfall	2
16	Heavy Snowfall	1
17	Rain Shower	9
18	Heavy Rain Shower	23
25	Thunderstorm	35
26	Heavy Thunderstorm	5
27	Storm	2

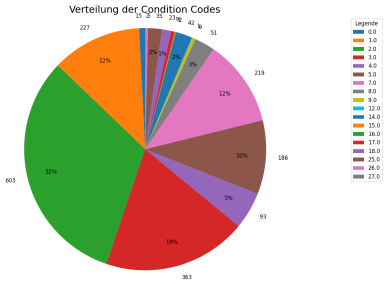

Abbildung 4: Verteilung der Condition Codes.

eingeteilten Condition Codes, ebenso eine höhere Anzahl an Ufos gesichtet wurden. Zum zweiten

6 FAZIT

Den Ergebnissen aus Kapitel 5 ist zu entnehmen, dass es, in Hinblick auf die verwendeten Vergleichdaten, bei vermeintlich besserem Wetter mehr Ufo-Sichtungen gemeldet werden als bei schlechterem Wetter. Da allerdings nur für eine sehr geringe Anzahl an Ufo-Sichtungen dazu passende Wetterdaten gefunden wurden, lässt sich die Aussagekraft der Ergebnisse nicht als endgültiges Ergebnis festlegen, sondern kann als Wegweiser für weitere Forschungen dienen. Eine weiterführende Herangehensweise wäre zum Beispiel die Betrachtung Wetterdaten von anderen Anbietern. Für Forschungen, welche außerhalb der Wetterzusammenhänge liegen, erweisen sich Uhrzeit der Sichtung (Morning Morality Effekt) sowie die Frage, ob es in der Nähe von Weltraumbahnhöfen, militärischen Einrichtungen oder Flughäfen zu überdurchschnittlich vielen Sichtungen kommt, als interessant. Ob es sich bei den vermeintlichen Ufo-Sichtungen wirklich um Ufos handelt, oder diese „Sichtungen“ nur Verwechselungen mit anderen Flugobjekten oder bewusste Fehlinformationen sind, können allein durch die verwendeten Daten nicht belegt werden.

REFERENZEN

[1] Peter Davenport. *The National Ufo Reporting Center*. 2021. URL: http://www.nuforc.org.

[2] Meteostat. *Meteostat Documentation*. 2021. URL: https://dev.meteostat.net.

[3] Meteostat. *Weather Condition Codes*. 2021. URL: https://dev.meteostat.net/formats.html#weather-condition-codes.

[4] NCEI. *API v1 Documentation*. 2021. URL: https://www.ncei.noaa.gov/support/access-data-service-api-user-documentation.

[5] NOAA. *API v2 Documentation*. 2021. URL: https://www.ncdc.noaa.gov/cdo-web/webservices/v2.

[6] Statista. *Sonnenstunden USA*. 2021. URL: https://de.statista.com/statistik/daten/studie/895421/umfrage/durchschnittlicher-anteil-der-sonnenstunden-in-staedten-in-den-usa-nach-monat/.

[7] L. Thomsen und U. Mammadov. *BigDataAnalytics Ufo*. 2021. URL: https://github.com/Lars0802/BigDataAnalyticsUfo.