**Introduction**

From our earlier [studies of UFO sightings](http://www.weirddatascience.net/index.php/2018/02/27/are-ufos-more-commonly-seen-near-us-military-bases/), a recurring question has been the extent to which the frequency of sightings of inexplicable otherworldly phenomena depends on the population of an area. Intuitively: where there are more people to catch a glimpse of the unknown, there will be more reports of alien visitors.

Is this hypothesis, however, true? Do UFO sightings closely follow population or are there other, less comforting, factors at work?

In this short series of posts, we will build a statistical model of UFO sighings in the United States, based on data [previously scraped](http://www.weirddatascience.net/blog/index.php/) from the [National UFO Reporting Centre](http://www.nuforc.org/) and see how well we can predict the rate of UFO sightings based on state population.

This series of posts is part tutorial and part exploration of a set of modelling tools and techniques. Specifically, we will use Generalized Linear Models (GLMs), Bayesian inference, and the [Stan](http://www.mc-stan.org/) probabilistic programming language to unveil the relationship between unsuspecting populations of US states and the dread sightings of extraterrestrial truth that they experience.

**Data**

As mentioned, we will rely on data from [NUFORC](http://www.nuforc.org/) for extraterrestrial sightings.

For population data, we can rely on the the [FRED](https://fred.stlouisfed.org/release?rid=118) database for historical US state-level census data. The combination of these datasets provides us with a count of UFO sightings per year for each state, and the population of that state in that year.

The downloading and scraping code is included here:

Show scraping code.

ZSH script to download via `curl`

#!/bin/zsh

# Download US state-level population datasets from FRED

# State series names are stored in the file 'series\_names' (downloaded from fred.stlouisfed.org)

#

#

# The per-series requests is included below.

export IFS=$'\n'

# Download

for state\_series in $(cat series\_names); do

curl -o "output/$state\_series.csv" "https://fred.stlouisfed.org/graph/fredgraph.csv?bgcolor=%23e1e9f0&chart\_type=line&drp=0&fo=open%20sans&graph\_bgcolor=%23ffffff&height=450&mode=fred&recession\_bars=on&txtcolor=%23444444&ts=12&tts=12&width=1168&nt=0&thu=0&trc=0&show\_legend=yes&show\_axis\_titles=yes&show\_tooltip=yes&id=$state\_series&scale=left&cosd=1900-01-01&coed=2018-01-01&line\_color=%234572a7&link\_values=false&line\_style=solid&mark\_type=none&mw=3&lw=2&ost=-99999&oet=99999&mma=0&fml=a&fq=Annual&fam=avg&fgst=lin&fgsnd=2009-06-01&line\_index=1&transformation=lin&vintage\_date=2019-03-04&revision\_date=2019-03-04&nd=1900-01-01"

done

Necessary ‘series\_names’ file:

WAPOP

GAPOP

CAPOP

MOPOP

DSPOP

ILPOP

TXPOP

NYPOP

FLPOP

ALPOP

COPOP

WIPOP

AZPOP

MIPOP

NCPOP

MAPOP

CTPOP

LAPOP

OHPOP

AKPOP

TNPOP

MNPOP

NJPOP

NMPOP

ARPOP

MDPOP

PAPOP

NVPOP

IAPOP

ORPOP

T5POP

DCPOP

HIPOP

NDPOP

KYPOP

VAPOP

IDPOP

KSPOP

INPOP

WVPOP

RIPOP

SCPOP

MSPOP

DEPOP

MTPOP

MEPOP

NEPOP

OKPOP

WYPOP

UTPOP

NHPOP

VTPOP

SDPOP

R code to combine data into tidy format

library( tidyverse )

# Read all CSV files

census\_files <- list.files( "output", full.names=TRUE )

# Join all data into a single table

census\_data <-

census\_files %>%

map( read\_csv ) %>% # Read each file, forming a list with an element for each

reduce( full\_join, by="DATE" ) %>% # Reduce (left to right) running a full join

dplyr::arrange( DATE ) %>% # Sort by date

gather( key="state", value="population", -DATE ) %>% # Gather to long format

transmute( date=DATE, state=str\_replace( state, "POP", "" ), population ) # Rename and tidy variables and names

# Output to an .rds

saveRDS( census\_data, "data/annual\_population.rds" )

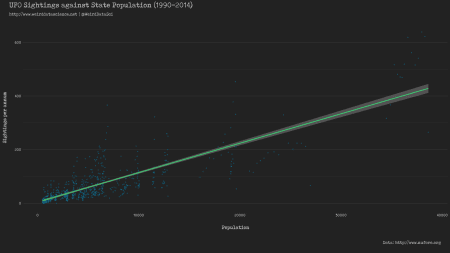
For ease, we will treat each year’s count of sightings as *independent* from the previous year’s — we do not make an assumption that the number of each sightings in each year is based on the number of sightings in the previous year, but is rather due to the unknowable schemes of alien minds. (If extraterrestrials visitors were colonising areas in secrecy rather than making sporadic visits, and thus being seen repeatedly, we might not want to make such a bold assumption.) Each annual count will be treated as an individual, independent data point relating population to count, with each observation tagged by state.

For simplicity, particularly in building later models, we will restrict ourselves to sightings post 1990, roughly reflecting a period in which the NUFORC data sees a significant increase in reporting and thus relies less on historical reports. (NUFORC’s phone hotline has existed since 1974, and its web form since 1998.)

**An Awful Simplicity**

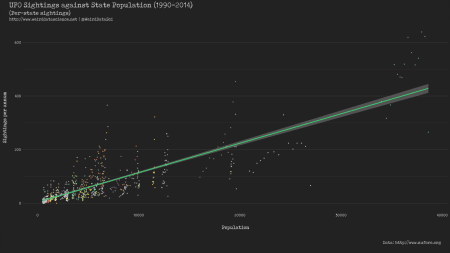
To begin, we start with the simplest form of model: a simple linear relationship between the count of sightings and the population of the state at that time. If sightings were purely dependent on population, it might be reasonable to assume that such a model would fit the data fairly well.

This relationship can be plotted with relative ease using the geom\_smooth() function of ggplot2 in R. For opening our eyes to the awful truth contained in the data, this is a useful first step.

[](https://i2.wp.com/www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-combined.png)

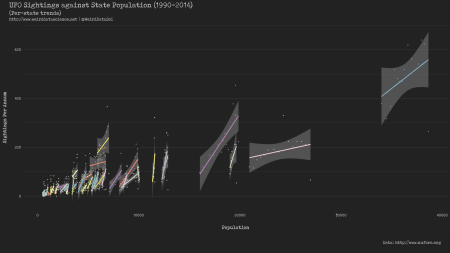
Global linear regression of UFO sightings against population. ([PDF Version](http://www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-combined.pdf))

While this graph does seem to support the argument that sightings increase with population *in general*, a closer inspection shows that the individual data points are clearly clustered. If we highlight the location of each data point, colouring points by US state, this becomes clearer:

[](https://i2.wp.com/www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-state.png)

Global linear regression of UFO sightings against population with per-state colours. ([PDF Version](http://www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-state.pdf))

This strongly suggests that, in preference to the simple linear model across all sightings, we might instead fit a linear model individually to each state:

[](https://i2.wp.com/www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-trends.png)

Per-state linear regression of UFO sightings against population. ([PDF Version](http://www.weirddatascience.net/wp-content/uploads/2019/04/lm_ufo_population_sightings-trends.pdf))

The code to produce the above graphs from the NUFORC and FRED data is given below:

Show data preparation and visualization code.

Prepare and combine datasets:

library( tidyverse )

library( magrittr )

library( lubridate )

# Prepare data for model fitting (and plotting)

# Load US population and UFO datasets

ufo <- read\_csv( "data/ufo\_spatial.csv" )

census <- readRDS( "data/annual\_population.rds" )

# Process UFO data to per-state counts per year.

# Drop Puerto Rico as we don't have census data. (Also, very few sightings -- 33 in dataset.)

ufo\_state\_annual <-

ufo %>%

# US only

filter( country == "us" ) %>%

# Apologies to Puerto Rico.

filter( state != "pr" ) %>%

# Convert date to year, drop all other variables except state.

transmute( date = year( as.POSIXct( datetime, format="%m/%d/%Y %H:%M" ) ), state=str\_to\_upper( state ) ) %>%

# Group by year

group\_by( date, state ) %>%

# Sum sightings

summarize( count = n() )

# Process census suitable for joining with UFO sightings.

# Drop "DS" state entry -- ("Department of State"?)

census <-

census %>%

filter( state != "DS" ) %>%

mutate( date=year( date ) )

# Join datasets

ufo\_population\_sightings <-

full\_join( ufo\_state\_annual, census )

# Missing data implies zero sightings.

# Restrict to post-1990 to avoid a high proportion of very small numbers of

# sightings.

ufo\_population\_sightings <-

ufo\_population\_sightings %>%

mutate( count = replace\_na( count, 0 ) ) %>%

filter( !is.na( population ) ) %>%

filter( date >= 1990 ) %>%

filter( date <= 2014 )

saveRDS( ufo\_population\_sightings, "work/ufo\_population\_sightings.rds" )

Fit linear trend in data via geom\_smooth() using a linear model.

library( tidyverse )

library( magrittr )

library( lubridate )

library( ggplot2 )

library( showtext )

library( RColorBrewer )

library( cowplot )

# Load UFO data

ufo\_population\_sightings <-

readRDS("work/ufo\_population\_sightings.rds")

# UFO reporting font

font\_add( "main\_font", "/usr/share/fonts/TTF/weird/Tox Typewriter.ttf")

showtext\_auto()

# Combined plot ignoring states.

ufo\_pop\_plot <-

ggplot( ufo\_population\_sightings, aes( x=population, y=count ) ) +

geom\_point( colour="#0b6788", size=0.6, alpha=0.8 ) +

geom\_smooth( method="lm", colour="#3cd070" ) + # UFO green

xlab( "Population" ) +

ylab( "Sightings per annum" ) +

theme\_weird() +

theme(

axis.title.y = element\_text( angle=90 )

)

# Construct full plot, with title and backdrop.

title <-

ggdraw() +

draw\_label("UFO Sightings against State Population (1990-2014)", fontfamily="main\_font", colour = "#cccccc", size=20, hjust=0, vjust=1, x=0.02, y=0.88) +

draw\_label("http://www.weirddatascience.net | @WeirdDataSci", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=0, vjust=1, x=0.02, y=0.40)

data\_label <- ggdraw() +

draw\_label("Data: http://www.nuforc.org", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=1, x=0.98 )

ufo\_pop\_titled <-

plot\_grid(title, ufo\_pop\_plot, data\_label, ncol=1, rel\_heights=c(0.1, 1, 0.1)) +

theme(

panel.background = element\_rect(fill = "#222222", colour = "#222222"),

plot.background = element\_rect(fill = "#222222", colour = "#222222"),

)

save\_plot("output/lm\_ufo\_population\_sightings-combined.pdf",

ufo\_pop\_titled,

base\_width = 16,

base\_height = 9,

base\_aspect\_ratio = 1.78 )

# Combined plot colouring states.

ufo\_pop\_plot\_states <-

ggplot( ufo\_population\_sightings, aes( x=population, y=count ) ) +

geom\_point( aes( colour=state ), size=0.6, alpha=0.8 ) +

geom\_smooth( method="lm", colour="#3cd070" ) + # UFO green

xlab( "Population" ) +

ylab( "Sightings per annum" ) +

scale\_colour\_manual( values=rep( brewer.pal( name="Set3", n=12 ), times=5 ) ) +

theme\_weird() +

theme(

axis.title.y = element\_text( angle=90 ),

legend.position="none"

)

# Construct full plot, with title and backdrop.

title <-

ggdraw() +

draw\_label("UFO Sightings against State Population (1990-2014)", fontfamily="main\_font", colour = "#cccccc", size=20, hjust=0, vjust=1, x=0.02, y=0.88) +

draw\_label("(Per-state sightings)", fontfamily="main\_font", colour = "#cccccc", size=16, hjust=0, vjust=1, x=0.02, y=0.48) +

draw\_label("http://www.weirddatascience.net | @WeirdDataSci", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=0, vjust=1, x=0.02, y=0.16)

data\_label <- ggdraw() +

draw\_label("Data: http://www.nuforc.org", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=1, x=0.98 )

ufo\_pop\_states\_titled <-

plot\_grid(title, ufo\_pop\_plot\_states, data\_label, ncol=1, rel\_heights=c(0.1, 1, 0.1)) +

theme(

panel.background = element\_rect(fill = "#222222", colour = "#222222"),

plot.background = element\_rect(fill = "#222222", colour = "#222222"),

)

save\_plot("output/lm\_ufo\_population\_sightings-state.pdf",

ufo\_pop\_states\_titled,

base\_width = 16,

base\_height = 9,

base\_aspect\_ratio = 1.78 )

# Combined plot colouring states with per-state trend lines.

ufo\_pop\_plot\_states\_trends <-

ggplot( ufo\_population\_sightings, aes( x=population, y=count ) ) +

geom\_point( aes( colour=state ), size=0.6, alpha=0.8 ) +

geom\_smooth( method="lm", aes( colour=state ) ) +

xlab( "Population" ) +

ylab( "Sightings Per Annum" ) +

scale\_colour\_manual( values=rep( brewer.pal( name="Set3", n=12 ), times=5 ) ) +

theme\_weird() +

theme(

axis.title.y = element\_text( angle=90 ),

legend.position="none"

)

# Construct full plot, with title and backdrop.

title <-

ggdraw() +

draw\_label("UFO Sightings against State Population (1990-2014)", fontfamily="main\_font", colour = "#cccccc", size=20, hjust=0, vjust=1, x=0.02, y=0.88) +

draw\_label("(Per-state trends)", fontfamily="main\_font", colour = "#cccccc", size=16, hjust=0, vjust=1, x=0.02, y=0.48) +

draw\_label("http://www.weirddatascience.net | @WeirdDataSci", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=0, vjust=1, x=0.02, y=0.16)

data\_label <- ggdraw() +

draw\_label("Data: http://www.nuforc.org", fontfamily="main\_font", colour = "#cccccc", size=12, hjust=1, x=0.98 )

ufo\_pop\_states\_trends\_titled <-

plot\_grid(title, ufo\_pop\_plot\_states\_trends, data\_label, ncol=1, rel\_heights=c(0.1, 1, 0.1)) +

theme(

panel.background = element\_rect(fill = "#222222", colour = "#222222"),

plot.background = element\_rect(fill = "#222222", colour = "#222222"),

)

save\_plot("output/lm\_ufo\_population\_sightings-trends.pdf",

ufo\_pop\_states\_trends\_titled,

base\_width = 16,

base\_height = 9,

base\_aspect\_ratio = 1.78 )

**Result**

The plots shown here strongly indicate that the rate of dread interplanetary visitations per capita varies differently per state. It seems, therefore, that while the number of sightings is generally proportional to population, the specific relationship is state-dependent.

This simple linear model is, however, entirely unsatisfactory in describing the data, despite its support for the argument that different states have different underlying rates of sightings.

In the next post, therefore, we will delve deeper into the unsettling relationships between UFO sightings and the innocent humans to which they are drawn. To do so, we will have to consider a class of techniques that go beyond the normal distribution that underpins key assumptions of the simple linear models used here, and so move into the eldritch world of *generalized linear models*.