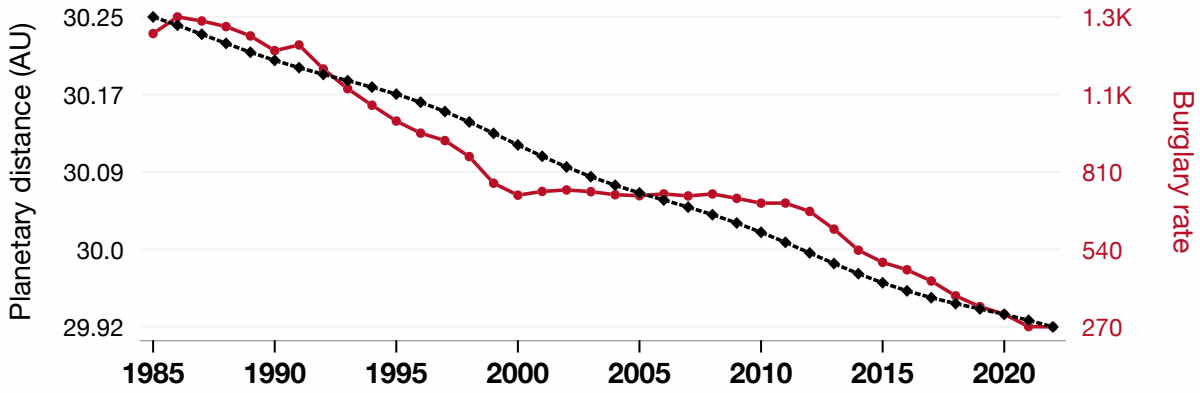
Spurious correlation #1,035 · View random

The distance between Neptune and the Sun

correlates with

Burglary rates in the US



The average distance between Neptune and the Sun as measured on the first day of each month · Source: Caclculated using Astropy

The burglary rate per 100,000 residents in United States · Source: FBI Criminal Justice Information Services

1985-2022, r=0.969, r^2=0.940, p<0.01 · tylervigen.com/spurious/correlation/1035

Download png, svg

AI explanation

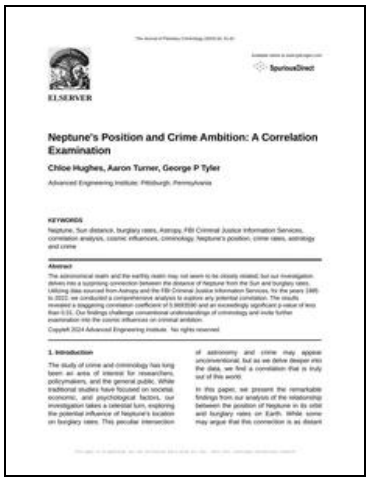
As Neptune cozied up to the Sun, its frigid temperatures and gaseous composition somehow emitted an anti-burglar force field. Criminals found themselves inexplicably drawn to the outskirts of the solar system, leaving Earth's neighborhoods surprisingly secure. It seems even the god of the sea has a knack for protecting earthly possessions.

Show image generation prompt

Show prompt for explanation

AI academic paper

(Because p < 0.01)



Show prompt used to generate this paper

Neptune's Position and Crime Ambition: A Correlation Examination

The Journal of Planetary Criminology

Jan 2024

Reminder: This paper is AI-generated. Not real!

Random correlation

Discover a new correlation

View all correlations

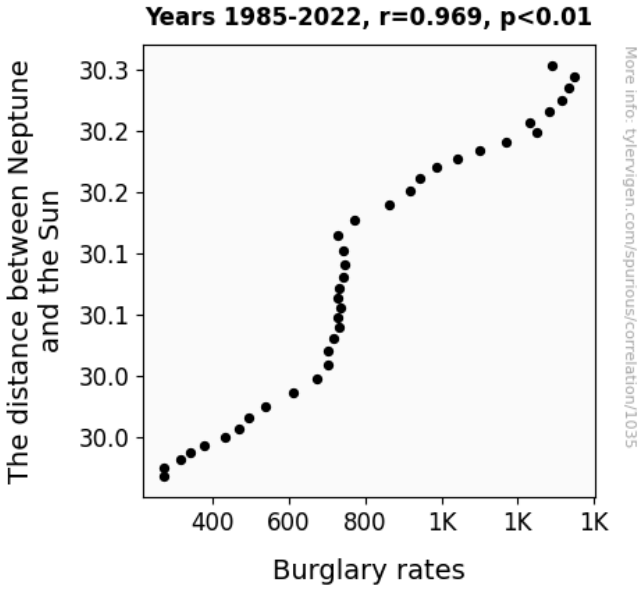
View all research papers

Report an error

Data details

The distance between Neptune and the Sun
Detailed data title: The average distance between Neptune and the Sun as measured on the first day of each month
Source: Caclculated using Astropy
Additional Info: *See full details*
*See what else correlates with **The distance between Neptune and the Sun***

Burglary rates in the US
Detailed data title: The burglary rate per 100,000 residents in United States
Source: FBI Criminal Justice Information Services
*See what else correlates with **Burglary rates in the US***



Correlation r = 0.9693590 (Pearson correlation coefficient)
Correlation is a measure of how much the variables move together. If it is 0.99, when one goes up the other goes up. If it is 0.02, the connection is very weak or non-existent. If it is -0.99, then when one goes up the other goes down. If it is 1.00, you probably messed up your correlation function.

r^2 = 0.9396569 (Coefficient of determination)
This means **94%** of the change in the one variable *(i.e., Burglary rates in the US)* is predictable based on the change in the other *(i.e., The distance between Neptune and the Sun)* over the 38 years from 1985 through 2022.

p < 0.01, which is statistically significant(Null hypothesis significance test)
The *p*-value is 1.5E-23.*Show* The *p*-value is a measure of how probable it is that we would randomly find a result this extreme.*Note* On average, you will find a correaltion as strong as 0.97 in 1.5E-21% of random cases. Said differently, if you correlated 65,334,709,717,884,718,809,088 random variables*Note* with the same 37 degrees of freedom, *Note* you would randomly expect to find a correlation as strong as this one.

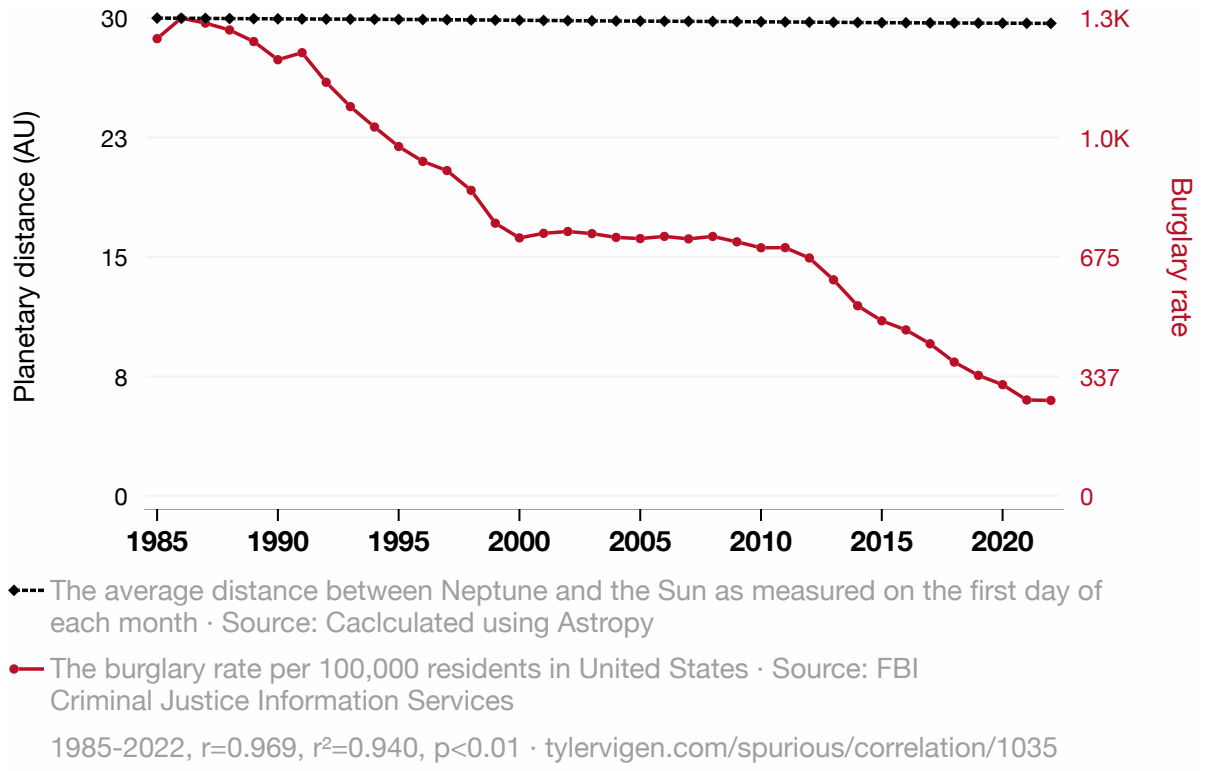
[0.94, 0.98] 95% correlation confidence interval (using the Fisher z-transformation)
Read more about the confidence interval

All values for the years included above: *Note*

	1985	1986	1987	1988	1989	1990	1991	1992	1
The distance between Neptune and the Sun (Planetary distance (AU))	30.2535	30.2445	30.2347	30.2248	30.2152	30.2064	30.1985	30.1912	30.1
Burglary rates in the US (Burglary rate)	1291.7	1349.8	1335.7	1316.2	1283.6	1232.2	1252.1	1168.4	109

Why this works

1. **Data dredging:** I have 25,153 variables in my database. I compare all these variables against each other to find ones that randomly match up. That's 632,673,409 correlation calculations! This is called “data dredging.” Instead of starting with a hypothesis and testing it, I instead abused the data to see what correlations shake out. It’s a dangerous way to go about analysis, because any sufficiently large dataset will yield strong correlations completely at random.
2. **Lack of causal connection:** There is probably[Note] no direct connection between these variables, despite what the AI says above. This is exacerbated by the fact that I used "Years" as the base variable. Lots of things happen in a year that are not related to each other! Most studies would use something like "one person" in stead of "one year" to be the "thing" studied.
3. **Observations not independent:** For many variables, sequential years are not independent of each other. If a population of people is continuously doing something every day, there is no reason to think they would suddenly *change* how they are doing that thing on January 1. A simple[Note] *p*-value calculation does not take this into account, so mathematically it appears less probable than it really is.
4. **Y-axis doesn't start at zero:** I truncated the Y-axes of the graph above. I also used a line graph, which makes the visual connection stand out more than it deserves. [Note] Mathematically what I showed is true, but it is intentionally misleading. Below is the same chart but with both Y-axes starting at zero.



Try it yourself

You can calculate the values on this page on your own! Try running the Python code to see the calculation results. *Show the steps to do this.*

```
# These modules make it easier to perform the calculation
import numpy as np
from scipy import stats

# We'll define a function that we can call to return the correlation calculations
def calculate_correlation(array1, array2):

    # Calculate Pearson correlation coefficient and p-value
    correlation, p_value = stats.pearsonr(array1, array2)

    # Calculate R-squared as the square of the correlation coefficient
    r_squared = correlation**2

    return correlation, r_squared, p_value

# These are the arrays for the variables shown on this page, but you can modify them to k
array_1 = np.array([30.2535,30.2445,30.2347,30.2248,30.2152,30.2064,30.1985,30.1912,30.18
array_2 = np.array([1291.7,1349.8,1335.7,1316.2,1283.6,1232.2,1252.1,1168.4,1099.7,1042.1
array_1_name = "The distance between Neptune and the Sun"
array_2_name = "Burglary rates in the US"

# Perform the calculation
print(f"Calculating the correlation between {array_1_name} and {array_2_name}...")
correlation, r_squared, p_value = calculate_correlation(array_1, array_2)

# Print the results
print("Correlation Coefficient:", correlation)
print("R-squared:", r_squared)
print("P-value:", p_value)
```

Reuseable content

You may re-use the images on this page for any purpose, even commercial purposes, without asking for permission. The only requirement is that you attribute Tyler Vigen. *Note* For more on re-use permissions, or to get a signed release form, see tylervigen.com/permission.

Download images for these variables:

- High resolution line chart *Note*
- High resolution line chart, optimized for mobile
- Alternative high resolution line chart
- Scatterplot
- Portable line chart (png)

- Portable line chart (png), optimized for mobile
- Line chart for only *The distance between Neptune and the Sun*
- Line chart for only *Burglary rates in the US*
- The spurious research paper: *Neptune's Position and Crime Ambition: A Correlation Examination*

View another random correlation

How fun was this correlation?

1 - Not for me | 2 - It is OK | 3 - Pretty good | 4 - It's great! | 5 - Awesome!

Correlation ID: 1035 · Black Variable ID: 1932 · Red Variable ID: 20084