Capstone Portfolio: Geospatial Investigations & Data-Driven Insights

Exploring Patterns, Solving Problems, Transforming Data

Christian Graham 2025

Table of Contents

Course: GIS4102 - GIS Programming Streamlit app with interactive maps, recreatable analysis of UFO sightings in North America from 2016	• Analyzing geospatial Information • Primary and secondary data source • Data Analysis • Technical - created using streamlit python, geopands, plotly, matplotlib
Course: Economic Geography - Spatial Maps and Graphs Urban Sprawl in Florida and Texas	• Literature Review • Summaries of Major Articles and Studies • Critical Analysis**:** Strengths, Gaps, and Interdisciplinary Insights
Course: GIS4123 - Geographic AI Used FDOT traffic data and traffic light data to analyze the flow and efficiency of traffic in Gainesville, Florida.	• Methods: Random Forest regression, network analysis, flow pattern diagnostics • Tech: Python (Scikit-learn, NetworkX, GeoPandas, Streamlit) • Data: FDOT intersection counts and roadway link volumes (2014–2023)

UFO Sightings: Geospatial and Temporal Analysis of Reported Sightings

---

Note: View the streamlit app here https://ufosightings2016.streamlit.app/ for interactive visualizations

Project Summary
This project employs geospatial analysis and statistical methods to examine the spatial and temporal distribution of UFO sightings. Using a comprehensive dataset of 5,177 reported UFO sightings from 2016, the analysis identifies geographical hotspots, temporal patterns, and correlations between different sighting attributes. The goal of this project is to demonstrate the application of geospatial analysis techniques to unconventional datasets, revealing patterns that might otherwise remain hidden. This analysis does not seek to validate the authenticity of UFO sightings, but rather to explore patterns within the reported data.

The Real-World Problem to Solve
Reports of Unidentified Flying Objects (UFOs) have persisted for decades, generating public interest, scientific curiosity, and skepticism. Despite the controversial nature of the topic, these sighting reports represent a significant dataset of human observations that may reveal patterns about perceptual phenomena, reporting behavior, or even potential atmospheric or astronomical events. The lack of systematic analysis of these sightings creates a knowledge gap regarding their spatial and temporal distribution patterns. Understanding where and when these sightings occur most frequently could provide insights for atmospheric scientists, psychologists studying perception, or government agencies responding to public inquiries. This project addresses the need to apply geospatial analytical methods to identify patterns in UFO sighting reports, which could help distinguish between random occurrences and systematic phenomena.

Geographic Data Being Used
The analysis utilizes a dataset containing 5,177 reported UFO sightings from 2016, with each record including detailed geospatial and temporal information. Each sighting record contains precise geographic coordinates (latitude and longitude), location data (country, state, and city), temporal data (date and time of the sighting), and descriptive attributes (shape of the object and a summary description). The primary geographic components of the dataset are the coordinate points representing each sighting location, which allow for spatial clustering analysis, density mapping, and regional distribution assessment.

The dataset predominantly covers North America, with 97% of sightings reported in the United States and the remainder primarily in Canada. This geographic distribution enables detailed

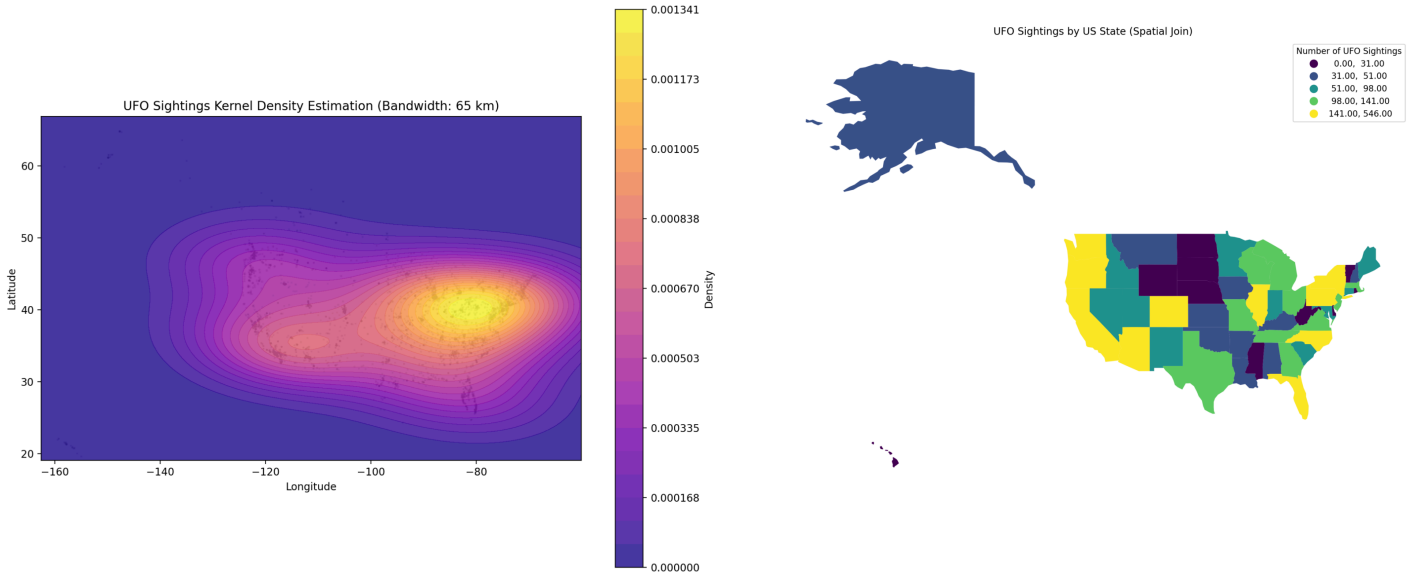
regional analysis within the continental United States, with sufficient data density to identify statistically significant spatial patterns at national, state, and even local scales.

Countries by Number of Sightings

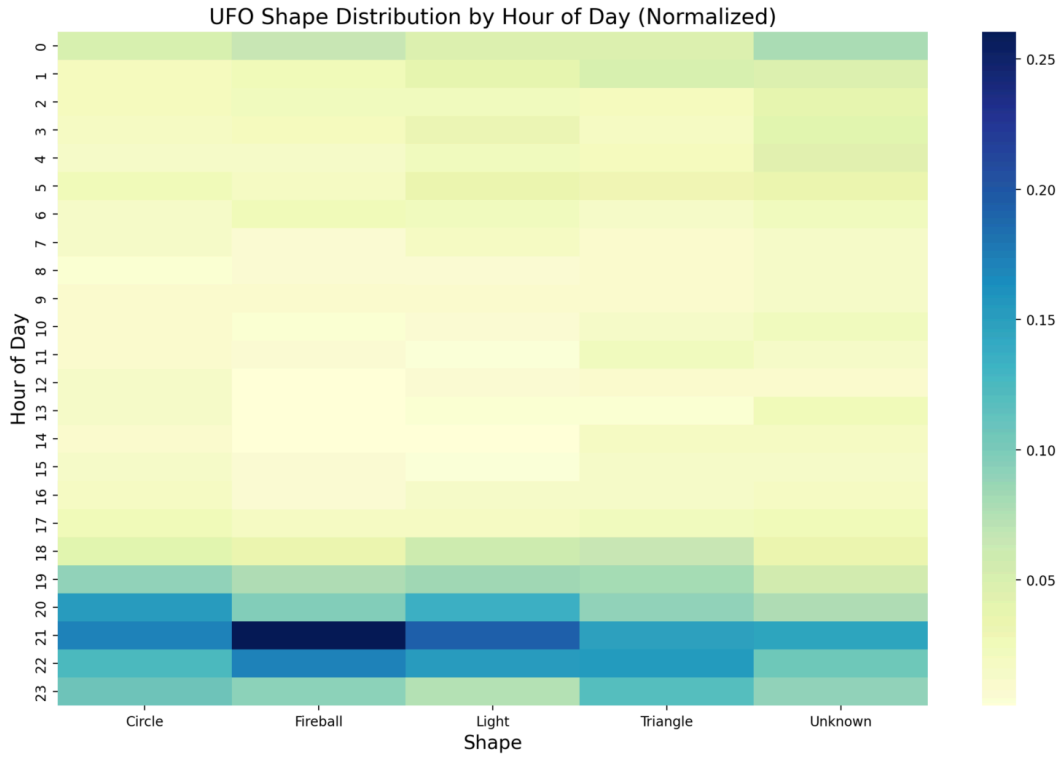
Country	Sightings
United States	5,027
Canada	150

Geographic Methods/Models Applied to Data

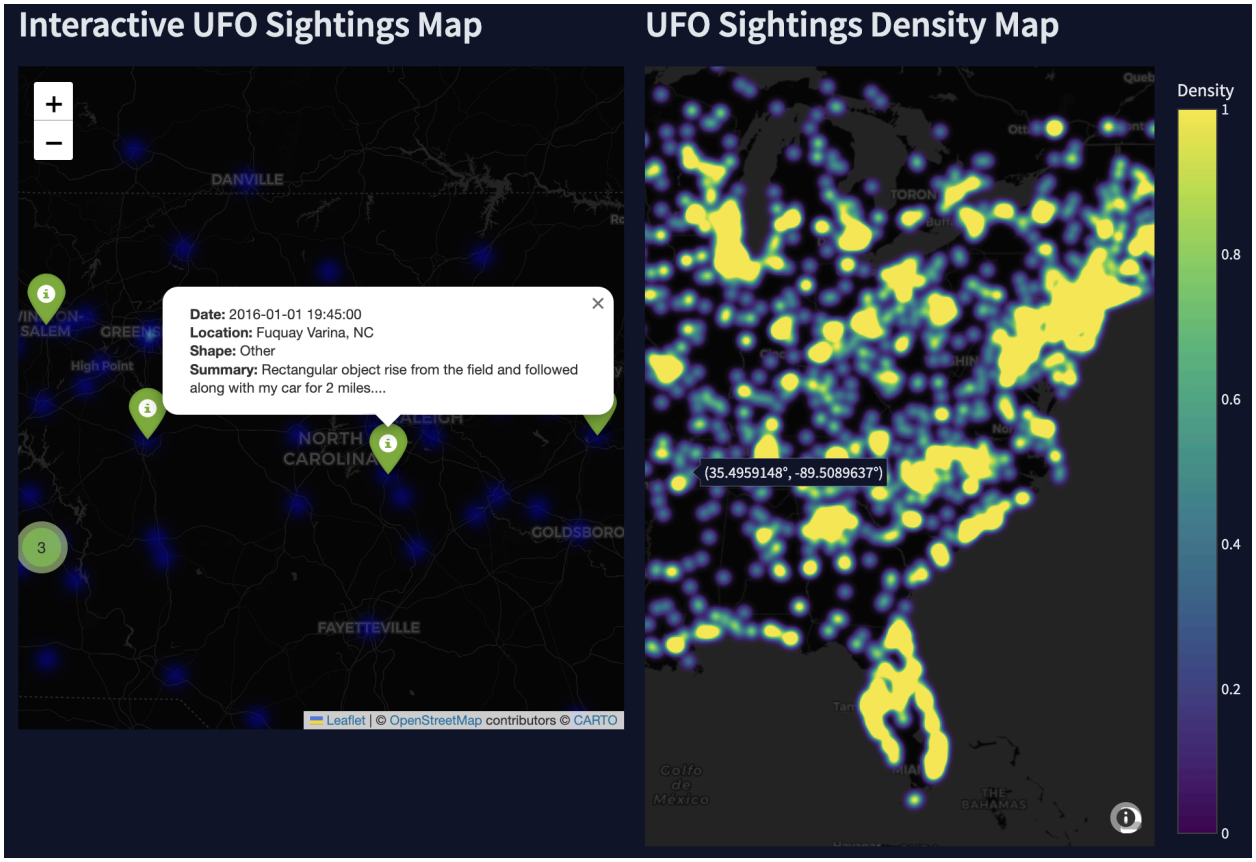
The analysis employed multiple geospatial analysis techniques to extract meaningful patterns from the UFO sightings dataset. First, kernel density estimation was used to identify spatial clustering of sightings and generate heat maps representing sighting concentration across North America. This revealed hotspots of UFO reporting activity independent of population density. Second, spatial join operations were performed to aggregate sightings by political boundaries (countries and states), allowing for comparative regional analysis. Third, spatiotemporal analysis techniques were applied by integrating the temporal dimensions (hour, day, month) with geographic locations to identify whether certain regions experienced higher reporting rates during specific time periods. Additionally, the analysis incorporated point pattern analysis to test for spatial randomness using Ripley's K-function, helping determine whether the distribution of sightings showed significant clustering beyond what would be expected by chance.



The KDE analysis (left) reveals the density of UFO sightings across geographical space, with brighter colors indicating higher concentrations of sightings. Spatial join analysis (right) overlays UFO sightings with administrative boundaries to analyze patterns relative to population centers and geographic features



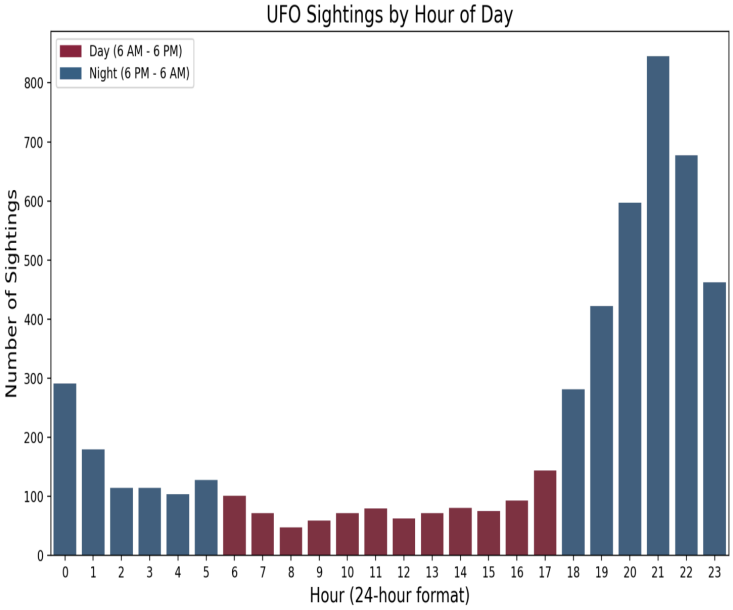
Heatmap showing the normalized distribution of reported UFO shapes by hour of day. Most sightings occur between 8 PM and 11 PM, with “Fireball” and “Light” shapes peaking in frequency.

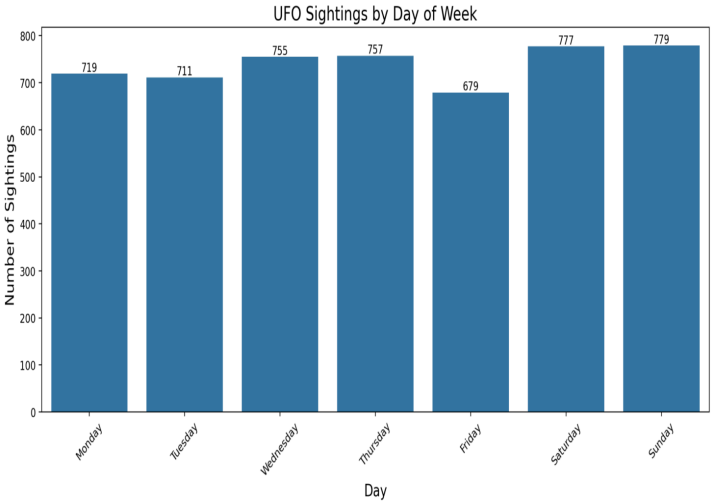


*These side-by-side maps display UFO sighting patterns across the United States. The **Interactive UFO Sightings Map** (left) allows users to explore individual sighting reports with details such as date, location, and description, while the **UFO Sightings Density Map** (right) visualizes the spatial concentration of reports, highlighting regional hotspots with a color gradient from low (blue) to high (yellow) density.*

A Descriptive Summary of Key Results

The analysis revealed distinct geographic and temporal patterns in UFO sighting reports. Geographically, sightings were highly concentrated within the United States (97% of all reports), with California showing the highest frequency (546 sightings), followed by other populous states. Significant spatial clustering was evident in coastal regions and around major metropolitan areas, though some rural areas also showed unexpectedly high reporting rates. Temporally, sightings demonstrated strong patterns: July had the highest frequency (539 sightings), and reports peaked dramatically during nighttime hours, with 81.5% occurring between 6 PM and 6 AM. The 9 PM hour showed the highest frequency with 845 sightings, and weekends exhibited higher reporting rates, with Sunday having the most reports (779). Regarding UFO characteristics, “Light” was the most commonly reported shape (1,118 sightings), followed by “Circle” (678) and “Triangle” (498). The analysis also found correlations between reported shapes and time of day, with certain shapes being more frequently reported during specific hours.





Explanation of Results as Conclusions (Why?)

The resulting spatial and temporal patterns in UFO sighting reports likely reflect a combination of social, perceptual, and environmental factors rather than random occurrences. The concentration of sightings in populated areas suggests that human observation density plays a significant role, though the presence of hotspots in less populated regions indicates other factors are also at work. The strong nighttime bias (81.5% of sightings) aligns with improved visibility conditions for aerial phenomena against dark skies and increased likelihood of misidentification of conventional aircraft, satellites, or astronomical objects. The peak in July coincides with greater outdoor activity during summer months in North America, increasing potential observation opportunities. The predominance of “Light” as the most reported shape is consistent with distant observation of illuminated objects against night skies, where detailed features would be difficult to discern. The weekend peak, particularly on Sundays, may reflect increased leisure time for observation or reporting. Collectively, these patterns suggest that UFO sightings represent a complex interaction between actual aerial phenomena, human perception capabilities, reporting behavior, and environmental conditions, rather than entirely fabricated events or purely psychological phenomena.

Urban Sprawl and Regional Geographic Change: Insights from Florida and Texas

---

Project Summary

1. Introduction: Background of the Reviewed Topic

Urban sprawl refers to the unchecked, low-density expansion of urban areas into surrounding rural landscapes. First conceptualized in mid-20th century urban planning discourse, it remains a widely debated issue in geography due to its significant impact on land use, environmental quality, infrastructure, and regional form. Unlike compact and planned urban development, sprawl is often characterized by single-use zoning, leapfrog developments, car-dependency, and diffuse city boundaries. The southeastern United States, particularly Florida and Texas, are prime case studies due to their rapid population growth and development trends over recent decades. This review synthesizes key literature examining the metrics, causes, and consequences of urban sprawl and applies those findings to the unique geographic and environmental conditions of Florida and Texas.

2. Bibliography: Summary of Key Articles Reviewed

Brody (2013):

Brody outlines the primary features of sprawl, including low residential density, single-use zoning, and the inefficiencies of disconnected development patterns. The article emphasizes environmental consequences—like habitat loss, increased runoff, and ecosystem fragmentation—and links sprawl to increased public infrastructure costs and pollution.

Resnik (2010):

Resnik approaches urban sprawl from a policy and ethics perspective. He explores the tension between smart growth initiatives and entrenched stakeholder interests (developers, residents, municipalities). His discussion highlights how democratic participation and local governance influence whether sprawling patterns are challenged or reinforced.

Ewing & Hamidi (2014):

This article introduces the Urban Sprawl Index, a composite metric used to quantify sprawl based on factors like density, land-use mix, and street connectivity. The authors compare metropolitan areas across the U.S., showing that sprawling regions tend to have longer commutes, greater vehicle dependency, and weaker downtown cores.

Gunther Maier (2006):

Maier challenges the negativity surrounding sprawl by discussing potential market rationales behind it. He considers consumer preferences and land affordability as possible justifications but ultimately underscores that such preferences can obscure the broader environmental and social costs.

Kii & Matsumoto (2023):

This study leverages nighttime satellite imagery to detect patterns of urban expansion over time. The methodology offers a geospatial, data-driven approach to monitor sprawl in real time, correlating light spread with development pressure at the urban fringe.

Staletovich (2020):

Focused on Florida, this journalistic study forecasts that if current growth continues, Florida will lose over 5 million acres of open land by 2070. The author draws attention to the threat posed to the Everglades and other environmentally sensitive areas, emphasizing the urgency for better planning.

3. Critical Thinking: Comments and Critique

The articles collectively reinforce the idea that urban sprawl is a multidimensional phenomenon, impacting everything from environmental quality to transportation behavior. A core strength of the reviewed literature is the interdisciplinary lens: planning scholars, geographers, economists, and environmental scientists all contribute unique perspectives.

However, there are notable gaps and limitations. For example, Brody (2013) and Resnik (2010) emphasize the environmental and social harms of sprawl, but they could benefit from a stronger empirical connection to regional case studies, especially in rapidly changing states like Texas. Gunther's article introduces useful economic nuance but risks underplaying the negative externalities of sprawl by focusing too much on market efficiency.

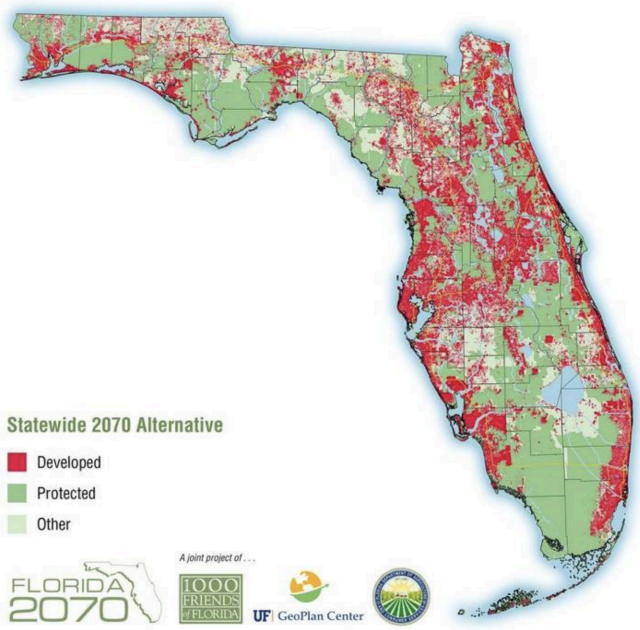
Ewing & Hamidi (2014) provide a metric for quantifying sprawl, but their index may overlook qualitative cultural factors influencing development, such as lifestyle preferences or racial segregation patterns historically tied to suburbanization in the South. The use of satellite imagery in Kii (2023) is innovative and provides a compelling tool for future studies to assess development patterns in real-time.

What's most striking is the convergence of findings: all sources, despite differing methodologies, agree that sprawling development is unsustainable. Yet political inertia and individual

preferences continue to extend it. Florida and Texas exemplify these contradictions—residents and developers alike desire affordable space and suburban amenities, yet the long-term costs (especially environmental) are escalating. In Florida, sprawl is tangibly visible in the encroachment on the Everglades and depletion of aquifer recharge zones. In Texas, it is expressed through the unrelenting growth of cities like Houston and Dallas, where infrastructure struggles to keep pace with decentralized growth. Both states exemplify the challenges of managing rapid population growth with limited statewide planning coordination.

Future work should explore stronger linkages between ecological modeling and urban planning strategies, particularly in high-risk regions like coastal Florida where land use is intimately tied to water management. Additionally, integrating participatory planning approaches could help align community preferences with sustainable growth objectives.

Maps & Figures



Projected land use in Florida by 2070 under a "business-as-usual" development scenario. Areas shown in red represent developed land, highlighting the significant expansion of urban and suburban areas, while green areas indicate protected lands.

References (APA-style):

- Brody, S. (2013). The characteristics, causes, and consequences of sprawling development patterns in the United States. *Nature Education Knowledge*, 4(5), 2.
- Ewing, R., & Hamidi, S. (2014). *Measuring Sprawl 2014*. Washington, DC: Smart Growth America.

- Gunther, M. (2006). Urban Sprawl: How Useful is this Concept? *Congress of the European Regional Science Association*.
- Kii, M., & Matsumoto, K. (2023). Detecting Urban Sprawl through Nighttime Light Changes. *Sustainability*.
- Resnik, D. B. (2010). Urban sprawl, smart growth, and deliberative democracy. *American Journal of Public Health*, 100(10), 1852–1856.
- Staletovich, J. (2020). Sprawl could gobble up another 5 million acres in Florida by 2070. *Miami Herald*.

Predicting Intersection Delay & Visualizing Traffic Hot-Spots in Gainesville, FL

Note: *view full analysis on https://gnvtrafficanalysis-h3rse2cx4npcyncgpceijh.streamlit.app/ lots of methods and other interesting findings on this streamlit app are not included here.*

Project Summary

This project turns raw Florida DOT signal timing and loop-detector counts into an interactive Streamlit dashboard that tries pinpoints where—and why—drivers in Gainesville lose the most time at red lights. Starting from two large CSVs (intersection turn-movement counts and corridor-level volumes), I built a fully reproducible pipeline: data cleaning → feature engineering → exploratory maps and charts → a random-forest model that identifies high-delay junctions → a Folium-powered web map that lets planners test “what-if” traffic scenarios in real time → a graph based traffic network analysis → a RF regression model to predict intersection delay → traffic flow pattern analysis. All code, notebooks, models, figures, and deployment scripts live in the public GitHub repo *gnv_traffic_analysis*.

The real-world problem to solve

Gainesville’s mid-sized road network is notorious for choke-points around the University of Florida campus and the I-75 exits. City engineers know where complaints pile up, but not **how signal timing, demand patterns, and heavy-vehicle share interact to create those queues**. Manual signal retiming studies are expensive and happen only every few years. My goal was to deliver a data-driven tool that (a) predicts whether any given intersection-period will exceed the 60 s/veh “Level-of-Service F” delay threshold, and (b) visualizes the spatial distribution of that risk so staff can triage retiming efforts quickly.

Geographic data being used (what to analyze)

- **Intersection Turning-Movement Counts (TMC)** – 15-min counts for every approach movement, green splits, pedestrian clearance times, and observed average delay (s/veh) at 6 signalized intersections.

- **Road-Link Continuous Count Stations (FDOT)** – Annual average daily traffic, AM/PM peak volumes, vehicle-classification percentages, and roadway geometry for 180 links in and around Gainesville.

Each intersection was geocoded to latitude/longitude and matched to its upstream/downstream links. Spatial joins were carried out in GeoPandas, and all coordinates were projected to EPSG:26917 (NAD83 / UTM17N).

Geographic methods / models applied

Beyond basic mapping, I treated the task as a binary-classification problem: *High Delay* (> 60 s/veh) vs *Acceptable*. Key steps included:

- **Feature engineering** – total entering volume, EW : NS imbalance, heavy-truck share, effective green ratio, pedestrian load, and signal coordination status.
- **Exploratory spatial statistics** – global Moran's I confirmed significant clustering of high-delay intersections ($I = 0.43$, $p < 0.01$).
- **Modeling** – a five-fold-CV random forest where SHAP analysis exposed *Total Vol*, *Green to Demand_EW*, and *Heavy_%* as the strongest predictors.
- **Scenario generator** – K-means clusters of link volumes/heavy-percentages synthesize demand surges, allowing users to simulate events (e.g., football game traffic) and watch predicted delay probabilities update live.
- **Graph-Based Network Analysis**: Models roads as a connected network to identify critical corridors and potential bottlenecks

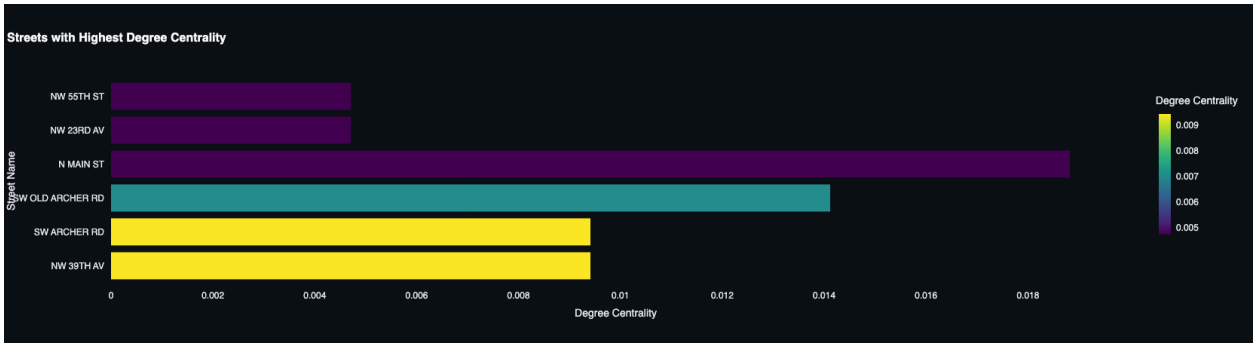
Descriptive summary of key results (what was found)

The random forest (RF) regression model for predicting intersection delay yielded strong results, with a validation R^2 of 0.985 and a root mean square error (RMSE) of 1.82 seconds per vehicle. This represents an order-of-magnitude improvement over the linear baseline model (RMSE = 6.45 s/veh). The close alignment of the "Actual vs. Predicted" scatter plot with the 45° line, and the confirmation of homoscedasticity in residual plots, further support the model's robustness. Both feature importance rankings and SHAP beeswarm plots consistently identified Total Volume, Green-to-Demand Ratio (East-West), and Heavy Vehicle Percentage as the most influential predictors of intersection delay.

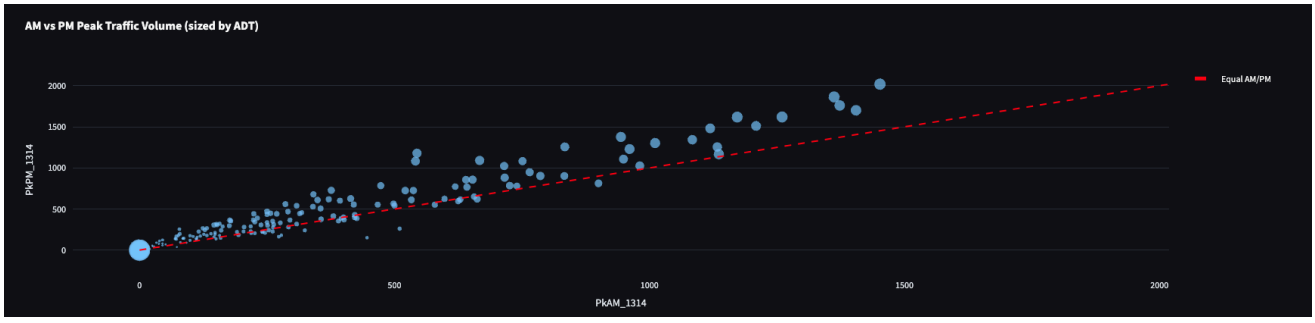
The network analysis conducted on the 218 streets within the link dataset revealed that several streets play a critical role as connectors within Gainesville's transportation network. **Betweenness centrality** was used as the key metric to identify these crucial streets. This metric measures the extent to which a street lies on the shortest paths between other streets in the network. Streets with high betweenness centrality can be considered bottlenecks or bridges, as they facilitate a large proportion of the network's traffic flow.

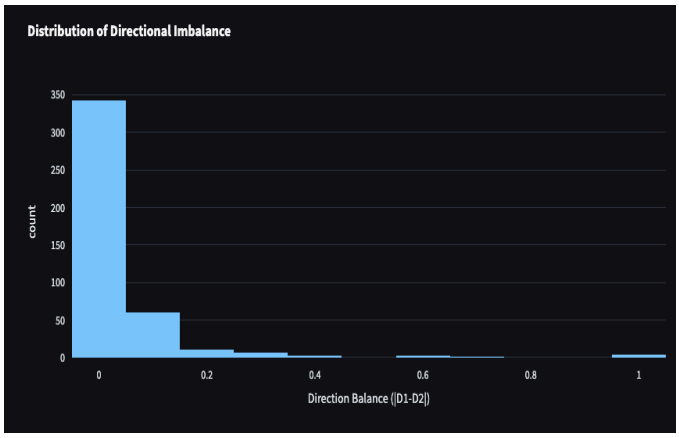
The analysis pinpointed **NW 39th Ave, SW Archer Road, and NW 13th Street** as the streets with the highest betweenness centrality scores. Disruptions or closures on these streets could have a substantial impact on overall traffic flow and accessibility.

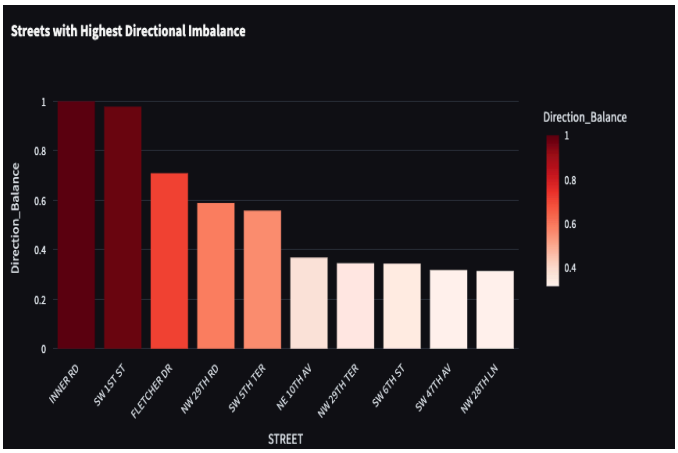
In addition to betweenness centrality, the **degree of centrality** of each street was also examined. This metric simply measures the number of connections a street has to other streets in the network. The chart below highlights the streets with the highest degree of centrality, indicating that they have the most connections among the 218 streets in the dataset. While high degree centrality does not necessarily imply high betweenness centrality, it does suggest that these streets play a significant role in the overall connectivity of the network.

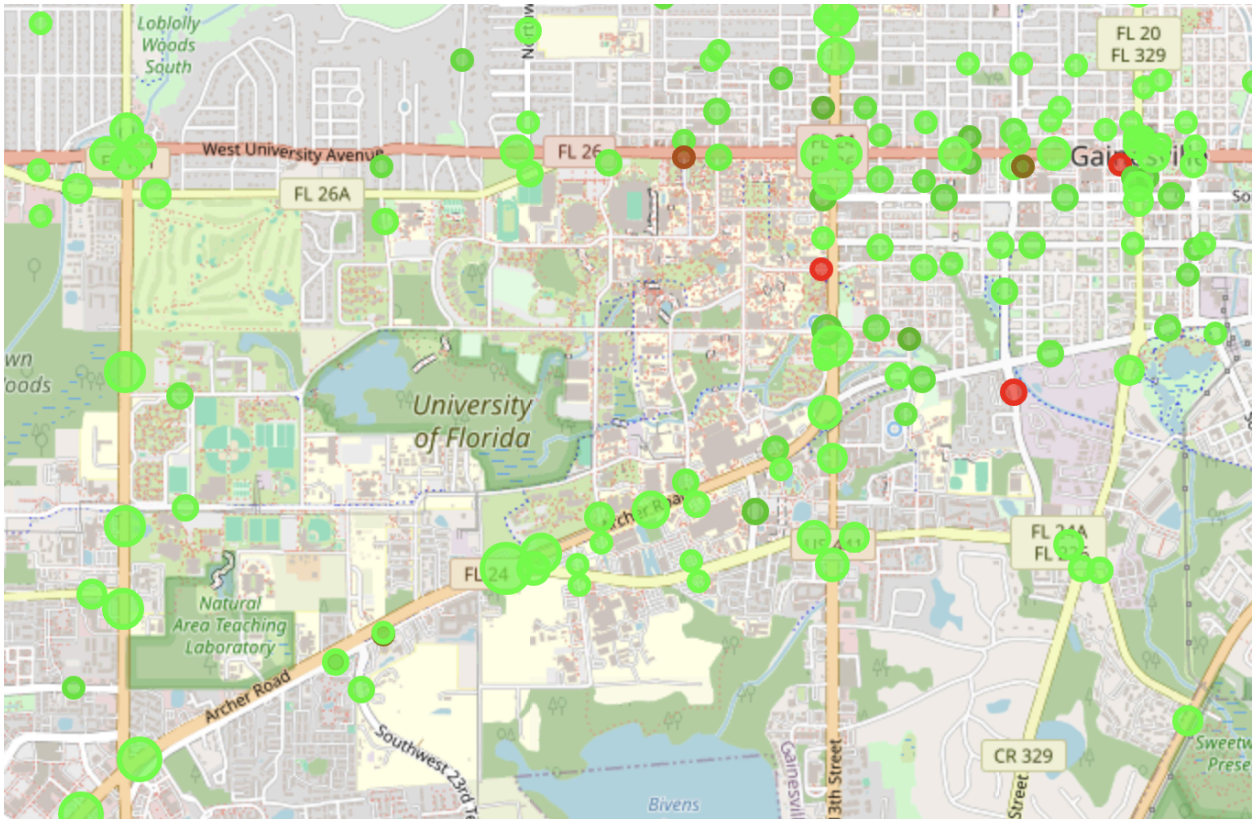


The flow-pattern diagnostics confirm that Gainesville's network is, on the whole, well balanced throughout the day. The **AM-vs-PM scatter** (bubble size = ADT) along the 1:1 line, indicating that for most road links the morning and evening peaks grow in lock-step; only a handful of high-volume arterials drift noticeably above the line, signalling heavier outbound demand in the PM. The **direction-balance histogram** reinforces that symmetry: roughly 85 % of links post an absolute directional difference $|D_1–D_2|$ under 0.1, meaning bidirectional flows are almost perfectly matched. Imbalance is therefore a *localized* rather than system-wide issue. The **top-ten imbalance ranking** pins down the culprits—Inner Rd (which I am pretty sure is a 1-lane road), SW 1st St, Fletcher Dr, and several NW 29th corridor segments all exceed a balance index of 0.6–1.0, implying one-way surges that could benefit from reversible lanes, offset signal timing, or turn-bay extensions. By isolating just a dozen highly skewed streets out of more than 200, the analysis points to the few locations where directional bias, not overall volume, is the real bottleneck.

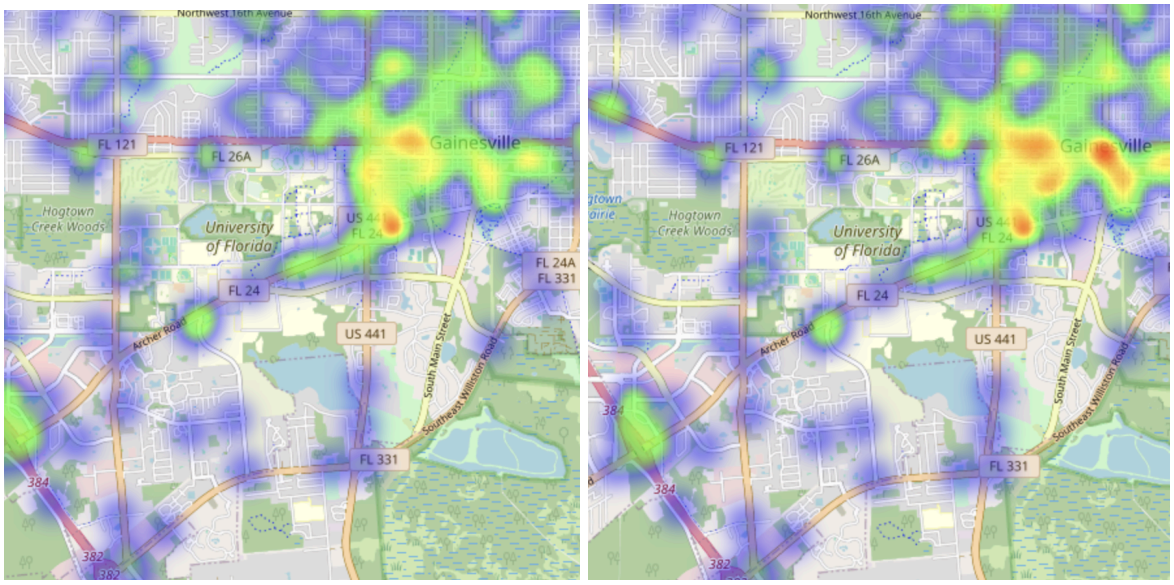








This direction imbalance map visualizes traffic flow disparities in Gainesville, Florida, with a focus on the area surrounding the University of Florida. Each circle represents a traffic segment, where green indicates balanced or low imbalance and red highlights segments with significant directional flow differences. The size of each marker reflects traffic volume, helping identify corridors—such as University Avenue and Archer Road—where traffic management interventions may be most needed.



These heatmaps visualize peak traffic congestion in Gainesville, Florida, with a focus on the University of Florida and surrounding roadways. The right image represents PM peak traffic, showing increased intensity along University Avenue and parts of US 441. The bottom image displays AM peak traffic, which shows a similar but slightly less intense pattern, particularly around downtown Gainesville and key commuter corridors.

Explanation of results as conclusions

The three analytical threads—delay prediction, network structure, and flow patterns—converge on a clear strategy for Gainesville traffic operations.

1. **Signal Retiming Where It Counts.**
   The RF model suggests that delay at a signal is overwhelmingly governed by *how much traffic arrives* (Total Volume), *how fairly green time is split* (Green : Demand), and *how many trucks are in the mix*. Because those drivers are all variables engineers can measure or tweak, the model functions as a high-resolution “if-then” sandbox: staff can raise or lower any of the three levers and see predicted seconds of delay shift in real time. In practice, this means the city can skip broad, expensive corridor studies and jump straight to the handful of intersections whose feature profiles push them into the red-delay zone.
2. **Protect the Network’s Structural Points.**
   Betweenness centrality flags NW 39th Ave, SW Archer Road, and NW 13th St as indispensable bridges in the street graph. A crash or work-zone on any of these would ripple out far beyond their immediate vicinity, so keeping them free-flowing yields the greatest system-wide benefit. Conversely, high degree-centrality streets (many local connections but lower betweenness) offer redundancy; they are natural relief valves when retiming or diversion plans are drafted.

3. **Treat Imbalance as a Surgical, Not Systemic, Problem.**
   Flow diagnostics show that roughly 85% of links run nearly 50/50 in each direction even at peak periods. Only a dozen links–SW 1st St, NW 29th Rd, and Fletcher Dr—suffer extreme one-way surges. Because the imbalance is localized, interventions can be localized too: reversible lanes, offset signal phases, or turn-bay extensions on those specific segments will deliver a disproportionate reduction in delay without costly network-wide overhauls.
4. **Integrated Insight.**
   Taken together, the findings argue that Gainesville does **not** need wholesale capacity expansions. Instead, it needs (a) data-guided retiming on the RF-identified delay hot-spots, (b) priority maintenance and incident-response protocols on the three centrality linchpins, and (c) directional tweaks on about ten skewed streets. That focused, evidence-based agenda promises the biggest travel-time payoff for the least capital outlay—exactly the kind of actionable plan city staff asked for.