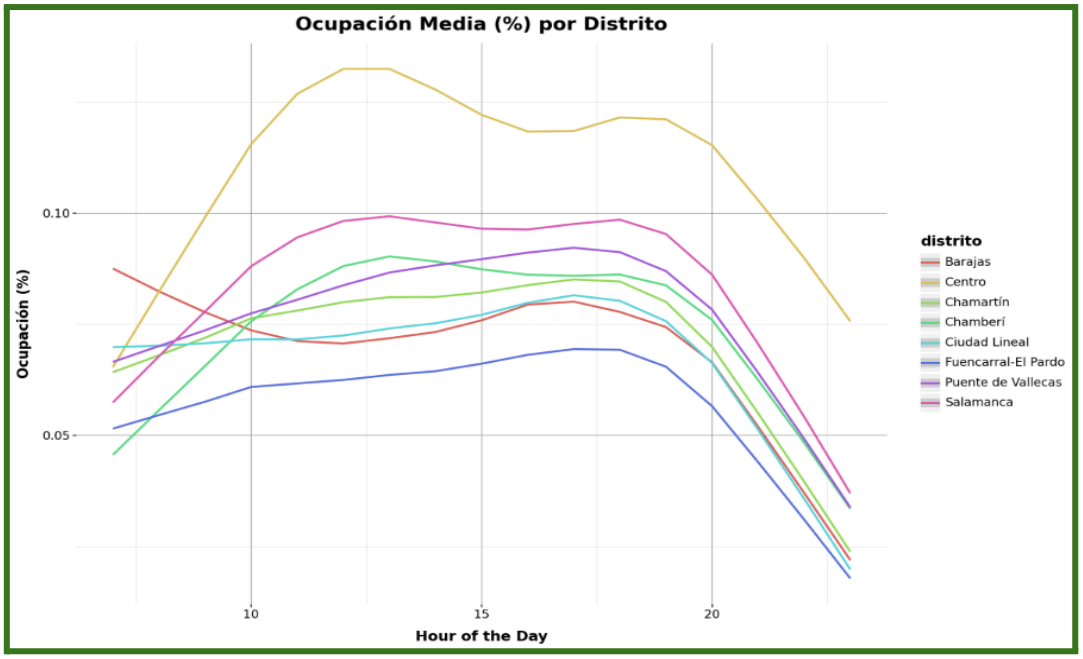
Luis Ignacio Pulido

Methodological Note

Traffic Congestion in Madrid

This project aims to analyze and predict the most and least favorable days to commute to work in Madrid based on urban traffic dynamics. The motivation behind the analysis is personal: I live approximately 1 hour and 30 minutes away from the office by public transport, and being able to anticipate high-congestion days could improve my daily planning, reduce stress, and optimize the use of remote work opportunities.

The core dataset was obtained from the Ayuntamiento de Madrid, consisting of minute-by-minute traffic occupancy data over a period of 90 days. Traffic occupancy (ocupación) is measured as a percentage between 0 and 100 and represents the proportion of time a traffic sensor is occupied by a vehicle. This makes it a strong proxy for real-time traffic saturation and flow. To contextualize this sensor data, I enriched it with geographical metadata — specifically, associating each traffic sensor with a Madrid district (e.g., Moncloa, Centro, Puente de Vallecas). This was made possible by cross-referencing a second dataset from the Ayuntamiento that links traffic sensor IDs to street segments and broader administrative zones, accounting for more than 30 million observations in total.



In order to capture external factors affecting traffic patterns, I merged the traffic dataset with weather data provided by the Agencia Estatal de Meteorología (AEMET). For each time interval, the dataset includes information on:

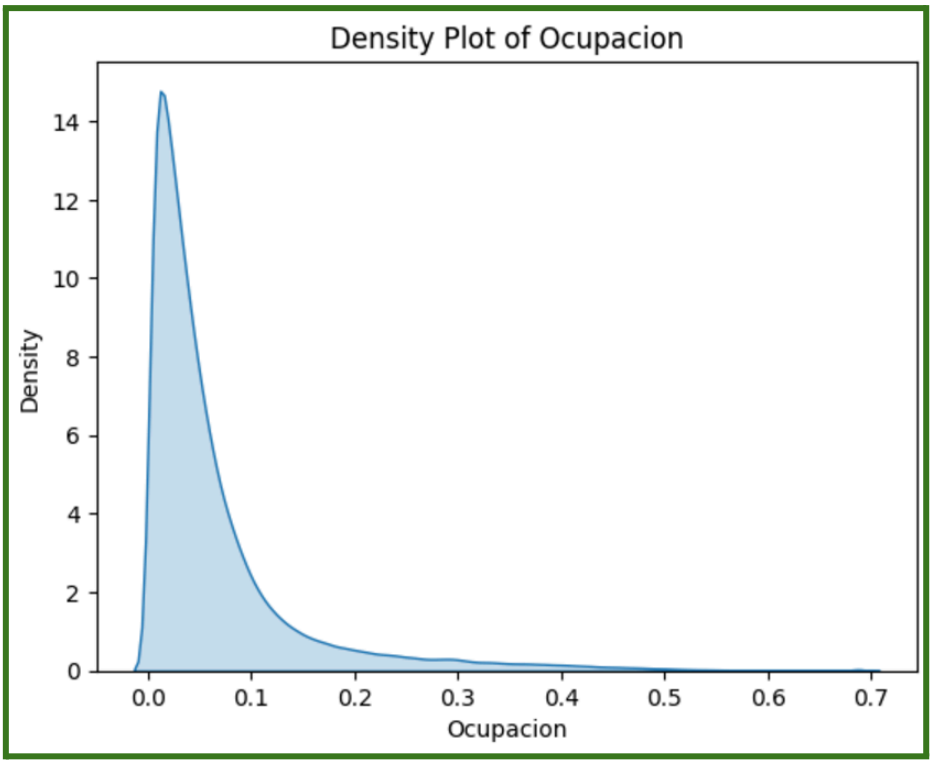
- Rainfall
- Wind intensity
- Temperature

Additionally, I incorporated temporal and contextual features such as:

- Day of the week
- Public holidays and adjacent "puentes"
- Hour of the day
- Upcoming holidays
- Whether there was a major event at the Santiago Bernabéu stadium

Underlying Distributions & Model Specification

Our target variable — traffic occupancy, measured as a percentage — is highly skewed, and in many cases behaves like count data (e.g., the number of "occupied moments" out of a possible interval). If we were to discretize or bin it (e.g., round to the nearest integer), we end up with something resembling counts per hour.



This justifies the use of a **Poisson approach**, which is designed to model the rate or count of events happening within a fixed time period and will give us the associated probability of observing *k* events in a fixed interval, given a known average rate λ (lambda), assuming the events are independent. We use the probability mass function (PMF) of the Poisson distribution:

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

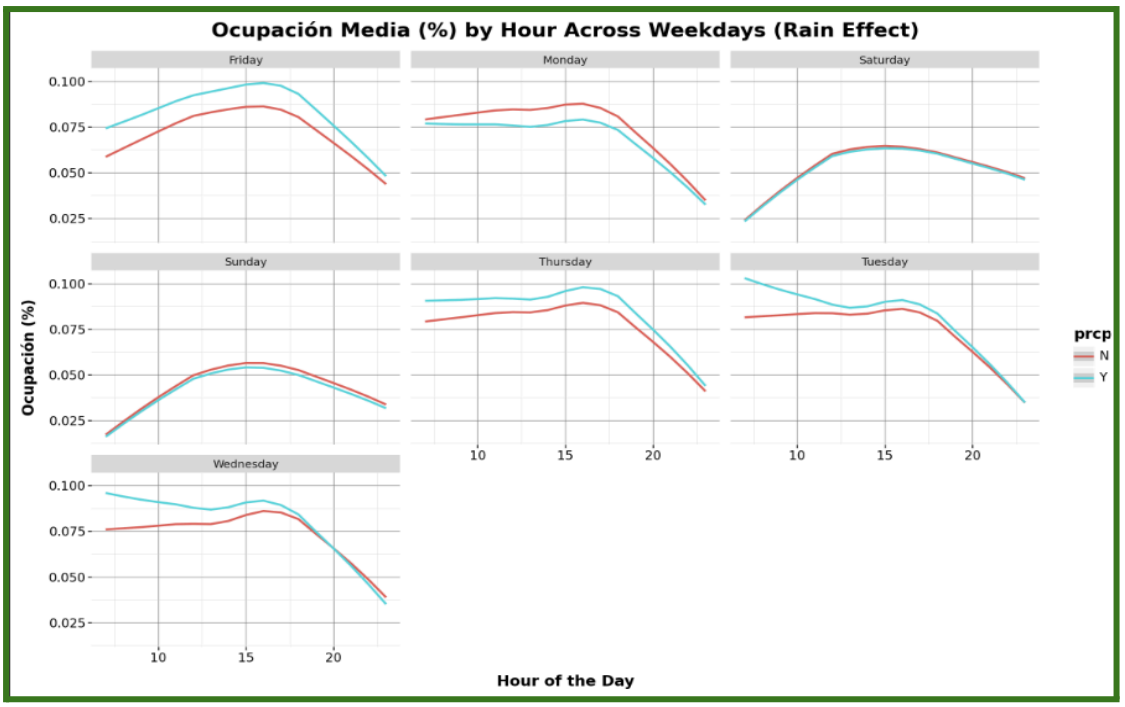
Initially, the modeling approach was based on this distribution, which assumes that the mean and variance of the response variable are equal. Although this seemed appropriate at first, as we were modeling the rate of traffic congestion as a function of time and context (e.g., day, weather), when fitting the model, we observed a clear case of overdispersion — the variance in the data was significantly higher than the mean, violating a key assumption of the underlying distribution. This phenomenon is common in real-world data, especially when there is unobserved heterogeneity, or when events are not purely random and independent.

To account for this, we turned to the **Negative Binomial distribution**, which is a generalization of the Poisson model by introducing an additional dispersion parameter allowing the variance to exceed the mean, specifically following a quadratic form. This quadratic relationship between mean and variance makes the Negative Binomial model more flexible for data where the variability is not constant but increases with the expected rate.

Whether using Poisson or Negative Binomial, the key modeling idea is that the rate of traffic congestion λ is not constant and it is, instead, assumed to vary depending on other contextual attributes. This approach moves from simply assuming that congestion follows a certain distribution, to building a predictive model where the parameter itself is **learned from the data**. It allows us to say, for example:

“On rainy Mondays in Puente de Vallecas, during morning rush hour, the expected congestion rate is significantly higher than average.”

By doing this, the structure of the underlying distribution (Poisson or Negative Binomial) is preserved, but the rate parameter is allowed to dynamically change based on relevant features such as weather conditions, district, hour of the day, and holidays.



Beta Specification

After exploring Poisson and Negative Binomial models — both tailored for count data — I realized that my target variable, traffic occupancy, is fundamentally a proportion bounded between 0 and 1 (or 0% to 100%), not a count. This makes it an excellent candidate for a model based on the Beta distribution, which is designed to describe random variables that take continuous values strictly between 0 and 1 and is a very flexible choice because it can model skewed distributions (left or right), but it can also represent symmetric or bell-shaped distributions. Such a flexibility makes Beta models suitable for a wide variety of data shapes — a major advantage over other distributions that assume symmetry (e.g., Gaussian) or integer values (e.g., Poisson). Following the previous approach in which we first derive a distribution and then we fit a model specification, the probability density function (PDF) of a Beta distribution with parameters α *and* β is:

$$F(y;\alpha,\beta) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}}\,dt$$

With

$$\Gamma(n) = (n-1)!$$

(Gamma Function)

The Beta distribution is used to model the behavior of a specific proportion or percentage of a process and it is particularly useful when the quantity of interest is a probability between 0 and 1, and especially when that probability is uncertain or subject to variation.

In many cases, this proportion can also serve as a parameter for another distribution!

For example…consider the case of a biased die. We want to estimate the true probability θ that the die lands on 6. Since we don't know θ exactly, we can model our belief or uncertainty about its value using a Beta distribution and, as we collect data (e.g., 3 sixes out of 10 rolls), we update the parameters of the Beta distribution to reflect the observed outcomes. This allows us to perform Bayesian inference, meaning we combine prior beliefs with observed evidence to arrive at a more informed estimate of the true probability.

As a final approach, we fitted a log-normal distribution to the data which delivered the best performance on the testing set in terms of correlation between predictions and true values for our dependent attribute.

"Before allowing a model to learn from the data, it's essential that we learn from the data ourselves — by understanding its distribution, variable behavior, and the nature of the problem."

Methodological Insights

- After testing several approaches, a log-normal regression ultimately provided the best fit for the data. This was due to two key reasons:
  (1) The traffic occupancy variable was highly skewed, and
  (2) The values never reached 1.0 (typically maxing out around 0.7), which made

binomial and beta models less appropriate.

- Despite taking advantage of **millions of observations**, model predictions still had notable error. This leaves space for improvement by adding **more months of data**, additional features (e.g. school schedules), or exploring **different model structures**.

- By modeling the process, rather than treating the outcome as fixed, we allowed key parameters like the mean or rate of congestion to vary based on features (e.g., weather, day of week, region). This flexibility supports richer inference, such as estimating expected occupancy under different scenarios, or simulating hypothetical conditions.

- Classical statistical models (e.g., log-normal regression) can outperform more complex alternatives when properly applied. It's often worth trying simple, interpretable models before reaching for more advanced algorithms.

- The importance of preprocessing, understanding the underlying distribution of your data, and conducting exploratory statistical analysis cannot be overstated. These steps set the foundation for effective modeling.

Empirical Findings

- **Worst traffic days**: Tuesdays and Wednesdays

- **Best days**: Saturday and Sunday (unsurprisingly)

- **Worst districts**: Centro, Ciudad Lineal, and Moncloa

- **Best districts**: Hortaleza and Villaverde

- **Peak congestion hours**: 9–10 AM and 7–8 PM on average

- **Time of day** showed a **strong nonlinear effect** on traffic patterns, more impactful than most other features.

- **Holidays**: Occupancy drops significantly, with a smaller dip on the **day before holidays**, likely due to early closures or people leaving the city.

- **Rain** increased occupancy by an estimated **2–5%**, though its effect was **weaker** than that of time-of-day patterns.

- On **non-event Bernabeu days**, traffic occupancy **declined gradually over the day**. However, on **Santiago Bernabéu event days**, occupancy started lower in the morning but showed a **steadily increasing trend**, likely peaking around event time.