Bayes + Trip-Time Residual Time Series

Francesco Balocco

February 17, 2026

1 What you will do

1. A **Bayesian** pickup/dropoff model built from raw counts.
2. A trip-time prediction model using pickup and dropoff zones, followed by residual autocorrelation diagnostics (ACF/PACF) and an AR fit.

2 Data and holdout split (concept)

Inputs

You need one NYC TLC Yellow Taxi dataset file (April 2025 is used in the course materials). The variables are:

- pickup zone ID: `PULocationID`
- dropoff zone ID: `DOLocationID`
- pickup timestamp: `tpep_pickup_datetime`
- dropoff timestamp: `tpep_dropoff_datetime` (only needed for the trip-time section)

Weekly time slot

We convert each pickup time into a weekly time slot index:

$$s = 24 \cdot \texttt{dow} + \texttt{hour} \in \{0, \ldots, 167\},$$

where `dow` is day-of-week (Mon=0, . . . , Sun=6) and `hour` is hour-of-day (0–23).

Train/test (holdout)

We use a time-based split: **train** on the first part of the month and **test** on the last 7 pickup dates (a full week, i.e. 168 hours).

3 Part A: Pickup/dropoff model

3.1 Core idea (counts → probabilities)

We estimate conditional probabilities directly from counts (maximum likelihood):

$$\widehat{P}(A \mid B) = \frac{\#(A, B)}{\#(B)}.$$

In plain language: "probability" here means "how often it happened in training".

3.2 Pickups: $\widehat{P}(\mathrm{PU} \mid s)$

For pickup zone z and weekly slot s:

$$\widehat{P}(\mathrm{PU} = z \mid s) = \frac{\#(\mathrm{PU} = z,\ s)}{\#(s)}.$$

Rationale: pickup patterns depend strongly on time (rush hours, weekends), so we condition on the weekly slot.

Practical fallback: if a slot has zero training trips (rare), use the global pickup distribution $\widehat{P}(\mathrm{PU} = z)$.

3.3 Dropoff destinations: $\widehat{P}(\mathrm{DO} \mid \mathrm{PU}, s)$ with backoff

For destination zone d, pickup zone z, and slot s:

$$\widehat{P}(\mathrm{DO} = d \mid \mathrm{PU} = z,\ s) = \frac{\#(\mathrm{DO} = d,\ \mathrm{PU} = z,\ s)}{\#(\mathrm{PU} = z,\ s)}.$$

Why backoff is needed: some (z, s) combinations may never occur in training, especially when you condition on both pickup zone and time. If (z, s) is unseen, we back off to a simpler conditional model:

$$\widehat{P}(\mathrm{DO} = d \mid \mathrm{PU} = z) = \frac{\#(\mathrm{DO} = d,\ \mathrm{PU} = z)}{\#(\mathrm{PU} = z)}.$$

If the pickup zone z itself is unseen, back off again to the global $\widehat{P}(\mathrm{DO} = d)$.

3.4 Holdout simulation (posterior predictive check)

We simulate the holdout week while keeping pickup times fixed:

1. For each test trip with slot s, sample $\widetilde{\mathrm{PU}} \sim \widehat{P}(\mathrm{PU} \mid s)$.
2. Given $(\widetilde{\mathrm{PU}}, s)$, sample $\widetilde{\mathrm{DO}} \sim \widehat{P}(\mathrm{DO} \mid \widetilde{\mathrm{PU}}, s)$, using the backoff rules above if needed.

What you compare (conceptually):

- **Pickups per zone by hour:** counts of trips by pickup hour × pickup zone.
- **Dropoff destinations by pickup hour:** counts of destination zones for trips that *start* in a given hour (because the model conditions on pickup time).

Important: this pickup/dropoff model does *not* include travel time, so it does not try to match dropoff times.

3.5 Optional: a simple accuracy metric (no numbers)

A common summary metric is mean absolute error (MAE) across hour×zone cells:

$$\mathrm{MAE} = \frac{1}{|\mathcal{H}||\mathcal{Z}|} \sum_{h \in \mathcal{H}} \sum_{z \in \mathcal{Z}} \left| \widehat{C}_{h,z} - C_{h,z} \right|,$$

where $C_{h,z}$ is the observed count and $\widehat{C}_{h,z}$ is the simulated count. You can report the formula and the interpretation ("average absolute gap per cell") without reporting its value.

4 Part B: trip time + residual autocorrelation + AR

4.1 Trip time definition

For each trip, define travel time in minutes:

$$T = \frac{\texttt{dropoff\_dt} - \texttt{pickup\_dt}}{60}.$$

Rationale for filtering: real data can contain errors and extreme outliers (e.g., negative durations). For a stable teaching example, it is common to keep trips within a reasonable duration range experiment with it.

4.2 Very simple predictor: mean by route with backoff

We build the simplest model that uses only pickup and dropoff zones:

$$\widehat{T} = \widehat{\mu}(z,d) \quad \text{where} \quad \widehat{\mu}(z,d) = \mathbb{E}[T \mid \mathrm{PU} = z,\ \mathrm{DO} = d]$$

estimated as the sample mean in the training data for each route (z, d).

Backoff rules: if a route (z, d) is unseen in training, back off to:

$$\mathbb{E}[T \mid \mathrm{PU} = z],$$

and if needed back off to the global mean $\mathbb{E}[T]$.

4.3 Residuals

Define the residual (prediction mistake) in minutes:

$$e_i = T_i - \widehat{T}_i.$$

Positive residual means the trip was longer than predicted; negative means shorter.

4.4 Make a regular hourly residual time series

AR models assume observations are equally spaced in time. Individual trips occur at irregular times, so we aggregate residuals by pickup hour in the holdout week:

$$r_h = \frac{1}{n_h} \sum_{i \in h} e_i,$$

where h indexes hours and n_h is the number of trips that start in hour h. This produces a length-168 series (one value per hour in the holdout week).

4.5 Correlogram: ACF vs PACF (what they mean)

ACF (autocorrelation function). The ACF at lag k measures how strongly r_h is correlated with r_{h-k}. In plain language, it answers:

> “If the residual is high now, is it also high k hours later?”

It reflects both direct effects and indirect effects through shorter lags.

PACF (partial autocorrelation function). The PACF at lag k measures the *direct* relationship between r_h and r_{h-k} *after removing what can already be explained by lags* $1, 2, \ldots, k-1$. In plain language:

> “Does the value k hours ago still matter once you account for the more recent hours?”

Why we look at both. ACF gives you a quick picture of how persistence decays with lag and whether there are seasonal patterns (e.g. daily cycles). PACF helps you see which specific lags are “directly important”, which is useful when choosing an AR structure.

4.6 AR model (concept)

An AR(p) model uses the dependence on the last p values,

For example, an AR(1) for the hourly residual series can be written as:

$$r_h = \phi_1 r_{h-1} + \varepsilon_h,$$

where ε_h is a noise term.

Rationale: if residuals are autocorrelated, then the simple route-mean model is missing time-varying effects (traffic, events, etc.). AR is a compact way to model leftover temporal structure in the residuals.