Graded Assignment: Data Analysis Project

Bayesian Statistics Specialization: Course 2, Techniques and Models

Tiago B. Lacerda

2025-05-11

Executive Summary

Introduction

In modern mobile networks (4G and 5G), operators face the challenge of allocating limited resources efficiently to maintain and enhance service quality. A critical performance metric in this context is the Excellent Consistent Quality (ECQ) test success rate at individual sites. ECQ assessments evaluate whether networks consistently support demanding applications such as video streaming, video calls, and gaming, ensuring a seamless user experience .

These tests are typically conducted on Android or iOS devices with embedded SDKs in applications, contingent upon user consent. They measure key performance indicators (KPIs) including download speed, upload speed, latency, jitter, packet loss, and time to first byte, aligning with thresholds recommended for various demanding applications .

However, the variability in the number of tests across sites—some reporting only a handful while others report hundreds due to natural user mobility—poses a significant challenge. Naïve "site-by-site" estimates can be misleading: small samples may produce extreme rates simply due to chance, and citywide averages can obscure localized underperformance.

To address this, we propose a three-level Bayesian hierarchical model that nests individual sites within municipalities and municipalities within ANFs. This framework facilitates the sharing of information across levels, yielding stable, data-driven estimates of each site's true ECQ success probability while properly quantifying uncertainty.

In this report, we begin by clearly defining the problem and the key question we aim to answer: Which sites have a true ECQ success probability significantly below the network average, and how can we rank them for targeted interventions, accounting for both data scarcity and local variability? We then detail the model structure, inference methodology, and decision-making strategy.

Problem Definition

Our network comprises multiple sites scattered across a city, each running a varying number of ECQ tests. Some sites may report as few as 5–20 tests in a given period, while others conduct several hundred. The core challenge is to identify which sites genuinely underperform in terms of ECQ success rate and thus prioritize network improvement investments, without being misled by the randomness inherent in small test counts.

Specific Question

> *Which sites have a true ECQ success probability significantly below the network average, and how can rank them for targeted interventions, accounting for both data scarcity and local variability?*

By formalizing this question, we set the stage for applying a hierarchical Bayes model that "borrows strength" across sites and municipalities, producing posterior distributions for each site's success probability. These posteriors underpin credible intervals and ranking metrics that guide robust, data-informed investment decisions.

Data

In this report, we analyze ECQ test data collected in October 2024 across Brazil's Northeast region, encompassing 8 ANFs. Each data point corresponds to a specific network site, identified by its unique ENDERECO_ID. For every site, we have recorded the total number of ECQ tests conducted and the number of successful tests (TESTES_ECQ_OK), indicating instances where the network met the stringent performance thresholds defined by the ECQ metric.

Below is a summary of the data we will be using in our analysis.

```
## tibble [958 x 6] (S3: tbl_df/tbl/data.frame)
##  $ group_id     : int [1:958] 101 103 93 106 104 107 95 96 94 97 ...
##  $ ANF          : chr [1:958] "83" "83" "83" "83" ...
##  $ MUNICIPIO    : chr [1:958] "ARACAGI" "ARARUNA" "AGUA BRANCA" "AREIAL" ...
##  $ ENDERECO_ID  : chr [1:958] "PBAAG_0001" "PBAAN_0001" "PBABW_0001" "PBAEA_0001" ...
##  $ TESTES_ECQ_OK: num [1:958] 9 27 12 10 80 0 19 5 26 0 ...
##  $ TESTES_ECQ   : num [1:958] 13 57 28 12 150 1 21 12 31 12 ...
```