Bayesian Workflow on Time to Reach Global Top 20 in Tennis

Alireza Honarvar & Reza Soumi

Aalto University

December 13, 2024

Aalto University
School of Science

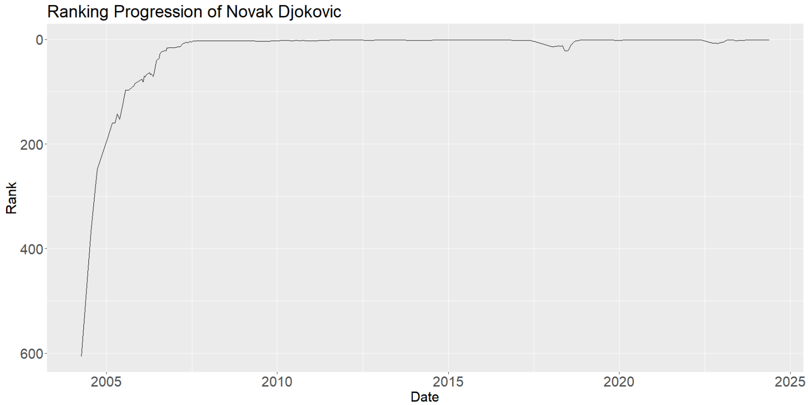
Overview

- Dataset: 106,716 rows containing 49 variables from Kaggle, such as:
  - **Match-Level**: Date, Surface, Duration
  - **Player Stats**: Player Name, Nationality, Age
  - **Rank Info**: Winner Rank, Loser Rank, Rank Points

Analysis Problem

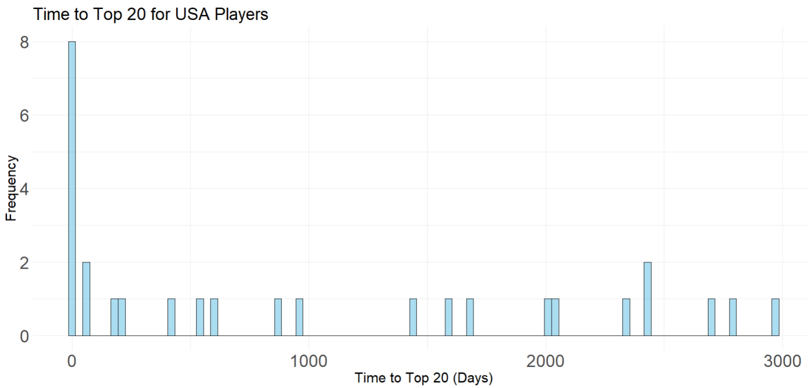
Aalto University School of Science

How long does it take players from different regions to reach the top 20?



Statistical Model: Gamma

Aalto University
School of Science

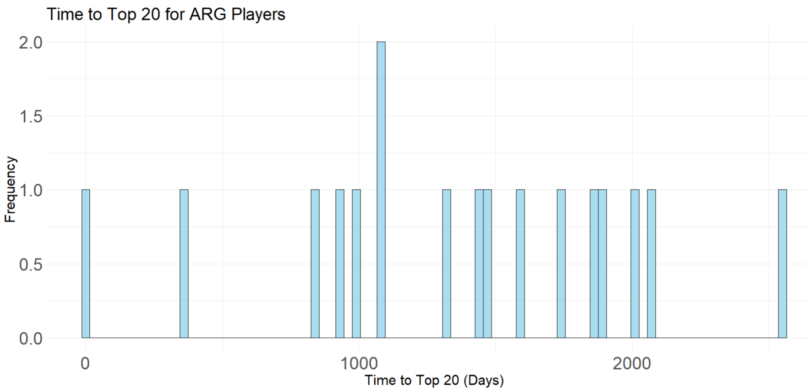


$$y_i \sim \text{Gamma}(\lambda_i, \theta), \quad \log(\lambda_i \theta) = \alpha + u_{\text{region}[i]} + u_{\text{country}[i]}$$

$$u_{\text{region}[i]} \sim \mathcal{N}(0, \sigma_{\text{region}}), \quad u_{\text{country}[i]} \sim \mathcal{N}(0, \sigma_{\text{country}})$$

Statistical Model: Gaussian

Aalto University
School of Science



$$y_i \sim \mathcal{N}(\mu_i, \sigma), \quad \mu_i = \alpha + u_{\text{region}[i]} + u_{\text{country}[i]}$$

$$u_{\text{region}[i]} \sim \mathcal{N}(0, \sigma_{\text{region}}), \quad u_{\text{country}[i]} \sim \mathcal{N}(0, \sigma_{\text{country}})$$

Weakly Informative Priors

- **Gaussian Model:**
  - Intercept: $\text{Normal}(1000, 350)$
  - Standard Deviations (Random Effects): $\text{Cauchy}(0, 2)$
- **Gamma Model:**
  - Intercept: $\text{Normal}(7, 1)$ (log-scale; approx. mean time of 1094 days)
  - Standard Deviations (Random Effects): $\text{Cauchy}(0, 0.5)$

Summary

- **Gaussian Model:**
  - Stable convergence, no divergences
  - Bulk ESS: 4959–8583, Tail ESS: 2059–2957
  - $\hat{R} = 1.00$
- **Gamma Model:**
  - Stable convergence, no divergences
  - Smaller effective sample sizes (Bulk ESS: 1346–4907, Tail ESS: 1346–3216)
  - $\hat{R} = 1.00$

Findings

- Gaussian model demonstrated more reliable and stable sampling

Posterior Predictive Checks

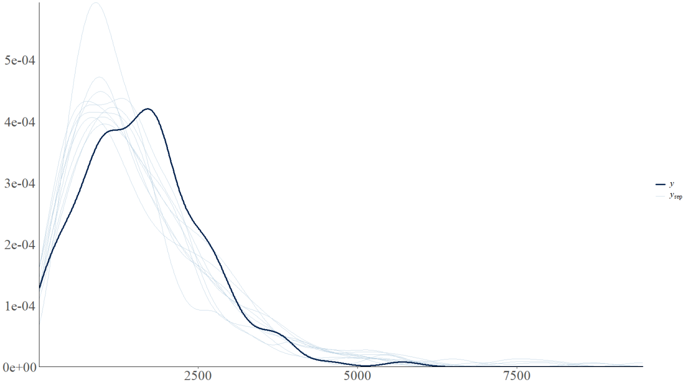


Figure: `pp_check` for Gamma

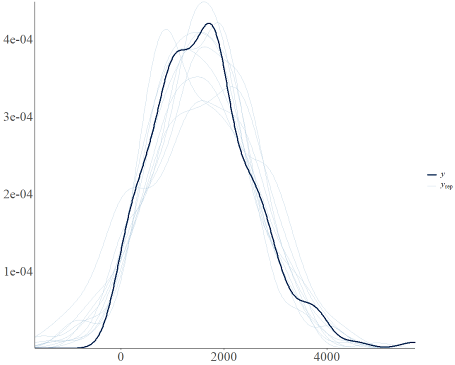


Figure: `pp_check` for Gaussian

Model Comparison

Aalto University
School of Science

Model Comparison (LOO-CV)

- **Gaussian Model:**
  - elpd_loo $= -1724.7$, SE = 12.0
- **Gamma Model:**
  - elpd_loo $= -1543.9$, SE = 8.9

Conclusion

- Gamma model performs better on fit and has better generalization.

Sensitivity Analysis

- **Gaussian Model:**
  - Shift $\mathcal{N}(1000, 1000) \to \mathcal{N}(10000, 10)$
  - elpd_loo $= -1726.6$, SE = 12.1
- **Gamma Model:**
  - Shift $\mathcal{N}(7, 1) \to \mathcal{N}(15, 1)$
  - Shift Cauchy$(0, 0.5) \to$ Cauchy$(0, 1)$
  - elpd_loo $= -1546.0$, SE = 8.9

Results and Discussion

Aalto University School of Science

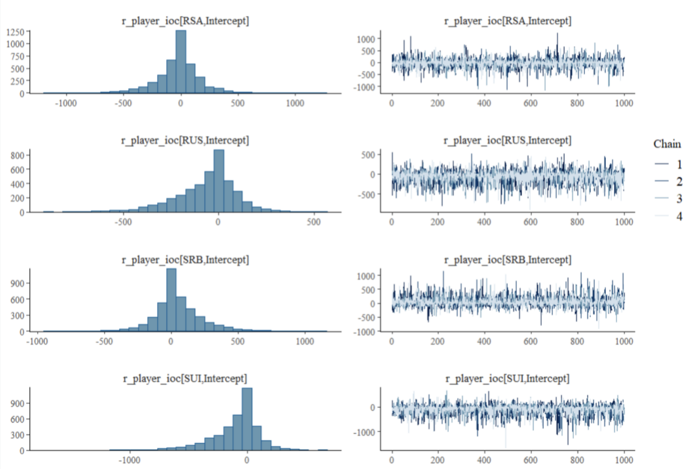


Figure: Player IOC and region effects on time taken to reach top 20

Results and Discussion

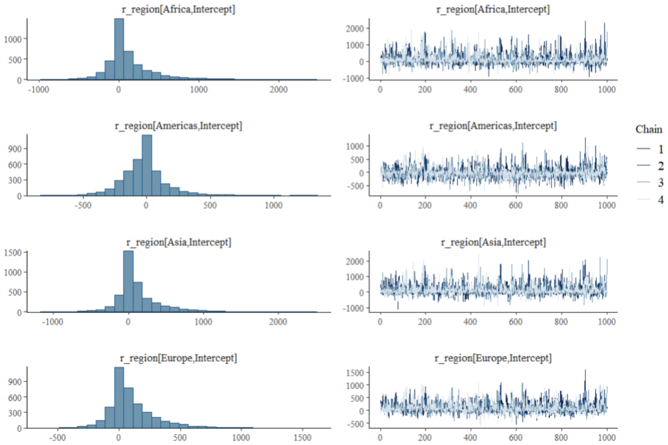


Figure: Player IOC effects on time taken to reach top 20

Aalto University
School of Science

Problem Definition

- Does region affect performance on different surfaces?

Model

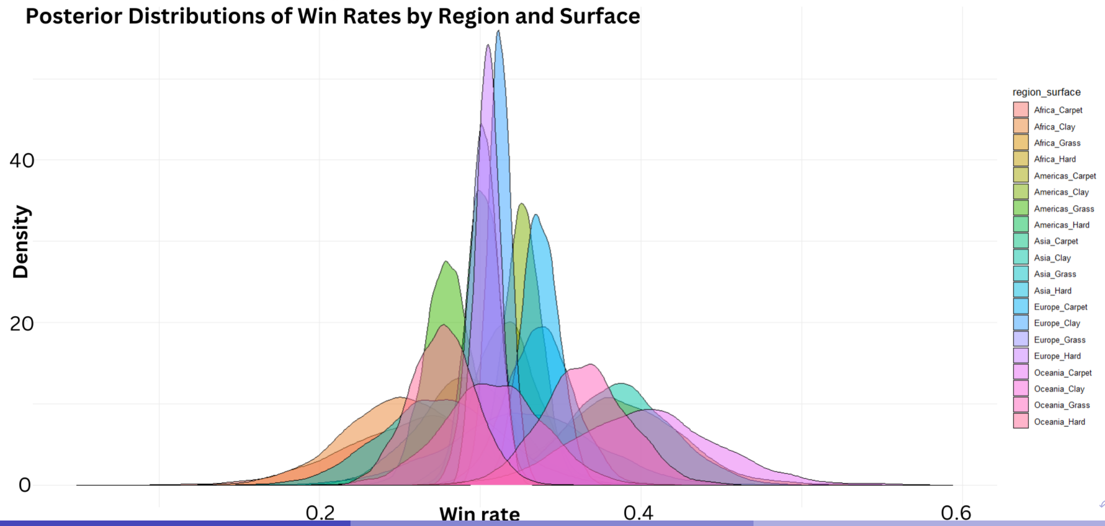
$$y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij})$$

Priors

$$\mu_{ij} \sim \mathcal{N}(0.35, 0.35), \quad \sigma_{ij} \sim \text{Cauchy}(0, 2)$$

(Extra) Results

Aalto University
School of Science



```
% GAUSSIAN
bayesian_model <- brm(
  time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc),
  data = time_to_top_20_regional_nonzero,
  family = gaussian(),
  prior = c(
    prior(normal(1000, 1000), class = "Intercept"),
    prior(cauchy(0, 3), class="sd", group="player_ioc"), # prior(normal(0, 1000)
        , class="sd", group="player_ioc")
    prior(cauchy(0, 3), class = "sd", group="region") # prior(normal(0, 1000),
        class = "sd", group="region")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  control = list(adapt_delta = 0.99, max_treedepth = 15)
)
```

```
% GAMMA
bayesian_model_gamma <- brm(
  time_to_top_20 ~ 1 + (1 | region) + (1 | player_ioc),
  data = time_to_top_20_regional_nonzero,
  family = Gamma(link = "log"), # Gamma distribution with log-link
  prior = c(
    prior(normal(1000, 1000), class = "Intercept"), # Approx. log(mean time
        around 1000)
    prior(cauchy(0, 2), class = "sd", group = "player_ioc"), # Prior on random
        effect std. dev
    prior(cauchy(0, 2), class = "sd", group = "region") # Prior on random effect
         std. dev
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  control = list(adapt_delta = 0.99, max_treedepth = 15)
)
```

```
% EXTRA
bayesian_models <- list()
for (i in 1:nrow(bayesian_data)) {
  region <- bayesian_data$region[i]
  surface <- bayesian_data$surface[i]
  win_rates <- bayesian_data$win_rates[[i]]
  bayesian_models[[paste(region, surface, sep = "_")]] <- brm(
    win_rates ~ 1,  # Single group model for win rates
    data = data.frame(win_rates = win_rates),
    family = gaussian(),
    prior = c(
      prior(normal(0.35, 0.35), class = "Intercept"),  # Centered around 0.5,
          wide SD
      prior(cauchy(0, 0.3), class = "sigma")  # Prior for standard deviation
    ),
    chains = 4,
    iter = 2000,
    cores = 4,
    seed = 123
  )
}
```