4 Methods

Yellow cards data set

The dataset consists of around 18,000 matches over a 5-year period across multiple European leagues including the top-two divisions in: England, France, Germany, Italy, and Spain.

- **fixture_id:** Identification number for the match.
- **season:** Season the match occurred in.
- **season_start_date:** Date the season started.
- **season_end_date:** Date the season ended.
- **country:** Country the match took place in.
- **league:** League the match took place in.
- **competition_level:** Binary whether the match was played in the 1st or 2nd division.
- **kick_off_datetime:** Date and time kick off took place in.
- **team1_name:** Name of team 1 – the home team.
- **team2_name:** Name of team 2 – the away team.
- **referee:** Referee who officiated the match.
- **sup_implied:** A translation of the available Asian handicap odds for these matches into what SmartOdds believed the market was expecting for (home) goals supremacy.
- **tg_implied:** A translation of the available Asian handicap odds for these matches into what SmartOdds believed the market was expecting for total goals in the match.
- **team1_yc:** Yellow cards received by team 1 – the home team.
- **team2_yc:** Yellow cards received by team 1 – the away team.
- **team1_rc:** Red cards received by team 1 – the home team.
- **team2_rc:** Red cards received by team 2 – the away team.
- **team1_bk:** Bookings/"points" accumulated by the home team according to the Pinnacle bookings rules laid out below.
- **team2_bk:** Bookings/"points" accumulated by the away team according to the Pinnacle bookings rules laid out below
- **bksup:** Final bookings supremacy score according to the Pinnacle bookings rules.
- **bktot:** Final total bookings count according to the Pinnacle bookings rules.

Pinnacle bookings rules

Each yellow card is worth one booking, and each red card is worth two Bookings. Second yellow cards on a competitor are ignored, so the maximum bookings for a player is three. Any cards shown to non-competitors (such as teammates on the bench, competitors leaving the pitch, the manager, coach, or other staff) are not counted. Cards shown during the half-time break are counted towards the 2nd-half period Bookings markets. Any cards shown after the whistle that ends regulation time will not be counted towards the markets for that game.

4.1 Bivariate Poisson distribution

The Bivariate Poisson distribution, as described by Karlis and Ntzoufras, is an extension of the Poisson distribution that is particularly useful when we are dealing with two related count variables. In this context, we consider three random variables Xk, where k=1,2,3. The joint distribution of the random variables X = X1 + X3 and Y = X2 + X3 then follows a Bivariate Poisson distribution, denoted as (X, Y) ~ BP (λ1, λ2, λ3), with a joint probability mass function given by:

$$PX,Y(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) = \exp\{-(\lambda 1 + \lambda 2 + \lambda 3)\} \ldots$$

Marginally, X and Y each follow a Poisson distribution with expected values E(X) = λ1 + λ3 and E(Y) = λ2 + λ3, respectively. The covariance between X and Y is given by λ3. Importantly, when λ3 = 0, X and Y are conditionally independent, and the Bivariate Poisson distribution simplifies to the product of two independent univariate Poisson distributions, known as the double-Poisson distribution. Therefore, λ3 not only quantifies the dependence between the two random variables, but it is the key parameter that differentiates the BP model from its more restrictive counterpart, the double-Poisson model (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003; Ley et al., 2019).

In conclusion, the Bivariate Poisson distribution, as detailed above, is uniquely equipped to capture any correlation that might exist between the two count variables, X and Y. This is particularly useful in the context of football matches, where the actions of one team can directly influence the actions of the opposing team. In our context, X and Y could symbolize goals scored or yellow cards issued by the home and away teams, respectively.

4.2 Bivariate Poisson modelling in football

In the context of football, the Bivariate Poisson distribution can be effectively used to model the number of goals scored by the home and away teams in a match. For instance, consider the goals scored by the home and away team in the ith game as (Xi, Yi). A BVP model for this scenario was proposed by Karlis and Ntzoufras (2003), where λ1i and λ2i represent the scoring rates for the home and away teams in the I-th game, respectively:

$$\begin{aligned} X_i, Y_i|\lambda_{1i}, \lambda_{2i}, \lambda_{3i} &\sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \\ log(\lambda_{1i}) &= \mu + \eta + att_{h_n} + def_{a_n} \\ log(\lambda_{2i}) &= \mu + att_{a_n} + def_{h_n} \\ log(\lambda_{1i}) &= \beta_0 + \gamma_1 \beta_{h_n}^{home} + \gamma_2 \beta_{a_n}^{away} + \gamma_3 \boldsymbol{\beta w_n} \end{aligned}$$

The scoring rates λ1i, λ2i are modelled as a function of several parameters. The parameter μ represents a constant intercept, while η captures the home advantage effect. Furthermore, the parameters $attt$ and def represent the attacking and defensive abilities of each team (t = 1, …, T), respectively, while the indices hn and an represent the home and away team playing in the nth game, respectively. In addition, the covariance parameter λ3 is modelled as function of several parameters. The parameter β_0 represents a constant intercept, while the parameters $\beta_{h_n}$and β_{a_n} represent the expected change in covariance, for the home team and away team respectively (Whittaker, 2011).

The model also incorporates a vector of covariates $\boldsymbol{wn}$, which are multiplied by a vector of coefficients $\boldsymbol{\beta}$. This allows for the inclusion of additional game-specific factors that might influence the scoring rates.

Moreover, the parameters $\gamma_1, \gamma_2, \gamma_3$ are dummy binary variables which activate different sources of the linear predictor for the covariance parameter $\lambda 3$ equation. For example, $\gamma_1 = \gamma_2 = \gamma_3 = 0$ we – the model considers the covariance to be constant, as in Egidi and Torelli (2020). Similarly, when $(\gamma_{1,}\gamma_{2,}\gamma_3) = (1,0,0)$ – the model considers covariance to depend solely on the home team, and so on. Karlis and Ntzoufras introduced parameters γ in such a way that it enhances flexibility and simplifies the process of testing various models.

The model above proposed by Karlis and Ntzoufras is useful for several reasons. First, the model captures fundamental factors that impact the number of goals scored in football matches; the model captures team effects, I.e., the attacking and defensive quality of the teams involved; the model captures the home advantage effect, I.e., the well-documented phenomenon of the home team generally having an advantage. In addition, the model supports the inclusion of covariates (that can influence the covariance between the two teams), which is essential to capture game-specific factors not already captured in the model parameters.

In conclusion, given the model's interpretability and demonstrated success (Ley et al., 2019; Egidi and Torelli, 2020), the model above provides a useful framework to build upon for the context of this dissertation, with respect to yellow cards. Namely, later in Chapter X, we develop a model for yellow cards inspired by the goals scored model described above.

4.3 Markov Chain Monte Carlo

- Estimation of parameters by MCMC (see next part)