Spatial Analysis of Fouls Across European Football Leagues

Name: Oliver Williams
Student Number: 240923037
Supervisor: Dr Fredrik Dahlqvist
Course: MSc Big Data Science

Abstract— This study investigates spatial and quantitative differences in foul behaviour across the top five European football leagues (Premier League, La Liga, Serie A, Bundesliga, and Ligue 1) during the 2015/16 season, using granular event-level data from the StatsBomb dataset. Across 1,823 matches, it was found there is significant variation in foul frequency: the English Premier League recorded the lowest average at 25.0 fouls per game, compared with 33.3 in Serie A and over 31 in La Liga and the Bundesliga.

Spatial analysis revealed that fouls are not uniformly distributed across the pitch, with league-specific hotspots. Using Gaussian Mixture Models, permutation testing, and Stouffer's Z method, France and Italy displayed statistically significant spatial divergence from all other leagues ($p < 0.01$), while Germany and Spain showed no overall difference ($p > 0.05$).

This research provides a statistically robust foundation for future studies on officiating, tactical adaptation, and league-wide consistency.

I. Introduction

Fouls are central to the structure and flow of football matches, influencing game tempo, player behaviour, and match outcomes. Despite their importance, the spatial and contextual dynamics of foul occurrences remain underexplored in empirical sports analytics. Most existing studies analyse fouls in aggregate, focusing on league-level trends or referee bias without considering the spatial distribution of fouls on the pitch.

This study addresses this gap by combining spatial statistics and modern machine learning techniques to analyse foul behaviour across Europe's five major football leagues during the 2015/16 season. Using high-resolution event-level data from the StatsBomb open dataset, this study examines foul frequency and location, and whether these patterns differ systematically between leagues.

Poisson regression is applied to analyse foul count differences and use Poisson point processes to model spatial intensity surfaces across the pitch. Gaussian Mixture Models (GMMs) are used to derive data-driven spatial zones, which serve as the basis for permutation testing of inter-league differences. Multiple testing is addressed through Holm's method, and overall spatial divergence is assessed using Stouffer's Z method to combine zone-level results.

II. BACKGROUND AND RELATED WORK

According to the official *Laws of the Game* set by FIFA, a foul occurs when a player commits an offence against an opponent that is penalised by the referee. Examples include kicking, tripping, holding, pushing, charging in a dangerous manner, or handling the ball deliberately. Some fouls lead to a *direct free kick* (e.g., tripping, holding, handball), while others result in an *indirect free kick* (e.g., dangerous play without contact, obstruction). Fouls inside the penalty area (henceforth referred to as the 'box') may result in penalty kicks, which are among the most decisive events in a match.

Fouls are subjective and can influence match flow, scoring opportunities, and disciplinary sanctions, and their interpretation has real consequences for teams and competitions. Subjectivity in refereeing can affect tactical approaches (e.g., defenders in Italy may play more cautiously in the box than in England), fairness in competition, and even commercial aspects like betting markets. By quantifying not only how many fouls occur but where they are most often given, this research highlights hidden biases and patterns in officiating that can benefit coaches, referees, analysts, league organisers, and commercial stakeholders.

A. Related literature

A substantial body of research has examined fouls and refereeing behaviour across Europe's top football leagues, often focusing on league-level comparisons and temporal trends. Sapp, Spangenburg, and Hagberg (2018), for example, analysed fouls and disciplinary actions in the Premier League, La Liga, Bundesliga, Serie A, and Ligue 1 over a ten-year period. They found a decline in fouls and cards per match since 2007–08, evidence of home bias, and league-specific tendencies—such as the English Premier League's higher aggression, measured by tackles per foul.

More recently, Nazarudin et al. (2024) used clustering and regression analyses on over 600 Big-5 league matches to profile referee strictness, confirming substantial variability in officiating styles. While valuable, these studies rely on aggregated statistics and rarely incorporate spatial or fine-grained temporal analysis.

Spatial statistical methods remain underused in football analytics. Foundational work by Baddeley et al. (2005) on Poisson point processes provides a rigorous framework for spatial event modelling, widely applied in ecology and epidemiology but seldom in sports.

The present study builds on this literature by combining Poisson point process regression, GMMs, permutation testing, and robust multiple testing corrections to analyse event-level foul data from several European leagues. This enables a detailed examination not only of how many fouls occur but precisely where they occur on the pitch, offering new insight into spatial heterogeneity in officiating across leagues, thus advancing beyond aggregate foul counts towards a more nuanced understanding of spatial dynamics in officiating.

B. AIC and BIC

When fitting statistical models, it is essential to determine the most appropriate model complexity. This decision is typically guided by model selection criteria that balance goodness-of-fit with simplicity. Two widely used methods for this purpose are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Both AIC and BIC assess how well a model fits the data while penalizing excessive complexity to avoid overfitting. However, they differ in how strongly they apply this penalty. AIC, rooted in information theory, is more focused on predictive performance and tends to favor models that fit the data well, even if they are more complex (Akaike, 1974).

In contrast, BIC, which is derived from Bayesian principles, applies a stricter penalty for the number of parameters, especially in larger datasets. This often results in the selection of simpler, more parsimonious models (Schwarz, 1978)

C. Poisson Regression and Negative Binomial model

Poisson regression is a widely used modelling approach for count data. Poisson regression assumes that the response variable follows a Poisson distribution and models the log of the expected count as a linear function of predictor variables (McCullagh and Nelder, 1989).

Poisson regression assumes that the mean and variance of the counts are equal. However, in practice, count data frequently exhibit overdispersion, where the variance exceeds the mean.

To address this, the Negative Binomial model is considered as an extension of the Poisson where appropriate, incorporating an additional parameter to account for overdispersion (Hilbe, 2011). This makes the Negative Binomial model more flexible and appropriate for analysing overdispersed count data, such as fouls in football matches.

D. Multiple Comparisons and the Holm Correction

In applied research, particularly when conducting numerous hypothesis tests, the risk of a Type I error (rejecting a true null hypothesis) increases with each additional test. This inflates the family-wise error rate (FWER), the probability of making at least one Type I error across all tests. To control for this, correction procedures are essential.

The Holm–Bonferroni procedure (Holm, 1979) is a widely used method for controlling the FWER. It offers greater statistical power than the traditional Bonferroni correction by sequentially adjusting p-values. This makes it particularly suitable for studies like this, where multiple group comparisons could otherwise lead to spurious findings.

E. Poisson Point Process

Poisson point processes provide a widely used statistical framework for modelling the spatial distribution of discrete events within a continuous domain. In this framework, events—such spatial occurrences in sports—are treated as points within a region. The main interest lies in determining whether these points are distributed randomly or show spatial structure.

The homogeneous Poisson process assumes a constant event intensity over space and independence between event locations. In contrast, the inhomogeneous Poisson process allows the intensity to vary as a function of spatial covariates or location. These models have found extensive applications in fields including ecology, epidemiology, and spatial econometrics (Baddeley et al., 2005).

While the primary use of Non-Homogeneous Poisson Point Process (NHPPP) models is often for descriptive statistical analysis, their utility as generative models for creating synthetic data is equally important. The thinning algorithm is a classic and widely used method for this purpose. A key paper that lays the groundwork for this approach is by Lewis and Shedler (1979), titled "Simulation of Nonhomogeneous Poisson Processes (NHPP) by Thinning." They present a simple yet robust method for simulating both one- and two-dimensional NHPPs by "thinning" a more easily generated homogeneous Poisson process. This approach is computationally efficient and conceptually straightforward, as it relies on a rejection-based sampling procedure. The method ensures that the final set of points perfectly matches the desired intensity function, making it a powerful tool for generating realistic spatial patterns.

The quadrat test, tests Complete Spatial Randomness (CSR). A foundational paper on this is "Modelling Spatial Patterns "Ripley (1977).

F. Gaussian Mixture Models (GMMs)

GMMs are a probabilistic approach commonly used to identify and describe complex clustering patterns in spatial data. They are particularly effective when the underlying distribution of events—such as events on a spatial field —is thought to arise from multiple overlapping sources or regions.

In spatial analysis, GMMs help capture spatial heterogeneity by modelling data as a mixture of simpler Gaussian distributions, each representing a cluster. Unlike hard clustering methods, GMMs support soft clustering, meaning each data point is assigned a probability of belonging to multiple clusters. This makes them especially useful in contexts like sports analytics, where player behaviour or foul patterns do not conform to rigid spatial boundaries and may overlap across zones or match situations (Rezek and Roberts, 2005; Bialkowski et al., 2014).

Parameter estimation in GMMs is typically performed using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), and the number of clusters is selected using model selection criteria such as the BIC or AIC.

G. Permutation Testing

To assess statistical significance, permutation testing is used by repeatedly reshuffling foul locations within each match or league context. This process generates a null distribution of spatial patterns that represent what would be expected if foul locations were randomly distributed. In each iteration, the data are randomly reassigned, a test statistic (e.g., spatial clustering or

distance-based measure) is calculated, and this is compared to the observed test statistic from the actual data. The key idea is that the more similar the two groups are in their spatial patterns, the more likely it is that a randomly generated test statistic will be more extreme than the observed. A statistically significant result occurs when the observed statistic is consistently more extreme than what would be expected under these random permutations.

H. Energy distance Test statistic

The Energy Distance is a statistical measure used to quantify the difference between two probability distributions by comparing the pairwise distances between sample observations. Unlike traditional statistical tests that assume specific distributional forms (e.g., normality), energy statistics are nonparametric and well-suited for high-dimensional or spatial data, where classical methods may fail to detect meaningful differences.

Originally developed by Székely and Rizzo (2004; 2013), the method has become widely used for two-sample testing, particularly in settings where distributional assumptions are difficult to justify. In spatial analysis, Energy Distance offers a flexible way to compare point patterns across groups, effective for analysing spatial distributions in various applications

Unlike binning- or kernel-based approaches, Energy Distance works directly with the geometry of the data, comparing distances within and between groups. This makes it especially useful in sports analytics, where event locations do not follow regular spatial structures and where preserving the spatial relationships between events is crucial (Aslan and Zech, 2005).

Pairwise Euclidean distances between event locations were computed using the cdist function from the SciPy library (Virtanen et al., 2020).

I. Stouffer's Z Method (Weighted)

Stouffer's Z method is a classical approach for combining p-values from multiple independent statistical tests into a single, overall measure of significance. The method converts each p-value into a z-score, allowing a weighted combination that accounts for differences in reliability or importance between tests.

The weighted version is particularly useful when individual tests vary in sample size or statistical power, enabling more informative aggregation by assigning greater influence to more reliable results. Under the null hypothesis and assuming independence of tests, the combined statistic follows a standard normal distribution and can be interpreted using a single p-value.

This method is applicable in various contexts for example, permutation test results from separate groups can be combined using Stouffer's method, with weights based on group size or number of observations (Whitlock, 2005).

J. Statsbomb Data

This study utilises the StatsBomb Open Data, a high-resolution public dataset of event-level football data. This granularity is essential for a spatial analysis, as it allows for the precise location of events on the pitch, a feature not present in traditional aggregate statistics (StatsBomb, n.d.). For this dissertation, the analysis focused on the 2015/16 season for the top five European leagues: La Liga, the Premier League, the Bundesliga, Ligue 1, and Serie A.

Compared to traditional aggregated match data, StatsBomb offers enhanced granularity that supports the application of advanced statistical techniques such as Poisson point process modelling and permutation-based inference. Previous studies have often relied on summary statistics or lower-resolution data, making this dataset particularly valuable for spatial modelling. Additionally, consistent pitch normalization and event labeling facilitate cross-league comparisons.

The specific events focused on are foul committed events. In the dataset foul committed events describe who committed a foul resulting in a free kick or a penalty.

III. METHODS

A. Statsbomb Data

The Statsbomb event data was taken for games played in the 2015/16 season across the top 5 European Leagues. In total this is 380 Premier League games, 377 Ligue 1 games, 306 Bundesliga games, 380 Serie A games and 380 La Liga games. All of these are full seasons except for Ligue 1 which is missing 3 games.

It contains each event that happens at a granular level with a categorized event type and appropriate related data. This paper is interested in the (x, y) pitch location data that allows identification of where fouls were committed on the pitch. While there are also corresponding events to show where the foul is won this doesn't include events such as handballs. In these cases a player can be noted as having conceded a foul but no opposing player would be noted as having won the foul so these corresponding events do not appear in the dataset.

$$x^* = |x - \max(x)| \quad (1)$$

$$y^* = |y - \max(y)| \quad (2)$$

To process, manipulate, and analyse the event-level data, this study employed the NumPy and pandas libraries. NumPy was used for high-performance numerical operations and the manipulation of multi-dimensional arrays (Harris et al., 2020), while pandas provided efficient tabular data structures and is widely used in statistical computing (McKinney, 2010). Match and event data were accessed via the statsbombpy package, which serves as a Python interface to the StatsBomb API, enabling direct and reproducible data retrieval (Pappalardo et al., 2019). For visualization, the matplotlib library was used for creating high-quality static, animated, and interactive plots (Hunter, 2007), and the mplsoccer package was used to plot match events and spatial distributions, providing flexible and publication-ready outputs tailored to football analytics (Wassenaar, 2021).

B. Poisson Regression

To assess whether foul frequencies differ across European leagues, a Poisson regression model was used. Let:

$$Y_i \sim \text{Poisson}(\lambda_{ij}) \quad (3)$$

represent the count of fouls in match i of league j, where:

$$log(\lambda_{ij}) = \beta_0 + \sum_{\{k=1\}}^{p} \beta_k X_{\{ijk\}} \quad (4)$$

Where:

- λij - represents the expected number of fouls occurring in a given match segment or spatial cell on the pitch, conditional on covariates.
- β_0 — The intercept term, representing the log-expected count when all predictors are zero.
- X_{ijk} — The k-th explanatory variable for observation i in match j.
- β_k — The regression coefficient for predictor X_{ijk}, interpreted as the log-multiplicative effect of a one-unit change in X_{ijk} on the expected event count, holding all other variables constant.
- p — The number of covariates in the model.

This model enables testing whether different leagues exhibit different behaviour.

Poisson and Negative Binomial regressions are implemented using the statsmodels Python package (Seabold and Perktold, 2010), specifically the statsmodels.api and statsmodels.formula.api modules.

There is a possibility the spatial areas being tested by can be reduced by remapping data along the data along the horizontal axis using this formula:

$$y' = \left| y - \frac{1}{2}\max(y) \right| \quad (5)$$

To assess whether foul distributions were symmetric about the horizontal midline, the pitch was divided into 5 vertically mirrored zones. Each of these mirrored zones were tested against each other to see if there was appropriate and consistent correlation.

Finally, a dispersion check was conducted to validate the Poisson assumption that the mean and variance are equal. The mean-to-variance ratio was computed as:

$$D = \frac{\mathrm{Var}(Y)}{E(Y)} \quad (6)$$

If D>1.5 or close to, indicating overdispersion, a negative binomial regression model was considered as an alternative.

C. Holm's Method for Multiple Testing Correction Regression

For each league, a hypothesis test was conducted to determine whether the mean foul count differed significantly from others. The resulting p-values were sorted in ascending order:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \quad (7)$$

Holm's method then applied a sequential testing procedure. For each ordered p-value p(i) , it was compared to an adjusted significance level:

$$p_{(i)} \leq \frac{\alpha}{m - i + 1} \quad (8)$$

Where α is the desired FWER (typically 0.05), and m is the total number of hypotheses tested.

This inequality is not a correction of the p-values themselves, but a decision rule: each $p(i)$ is evaluated against its threshold. If $p(i)$is greater than the threshold, the procedure stops, and all remaining hypotheses are retained. Otherwise, the null hypothesis is rejected, and the procedure continues.

The Holm Correction was implemented using the statsmodels Python package (Seabold and Perktold, 2010), specifically the statsmodels.stats.multitest modules.

D. Poisson Point Processes

To capture the spatial pattern of fouls within matches, foul locations were modelled using the Poisson point process. Each foul was represented as a two-dimensional point (x, y) in pitch coordinates, with all fouls from a match forming a realization of the process.

$$\lambda(x, y) = f(x, y) \quad (9)$$

where $f(x, y)$ is the spatial intensity function estimated non-parametrically using kernel density estimation (KDE).

KDEs were implemented using the Kernel Density class from the sklearn.neighbors module in Python, with a Gaussian kernel. This approach smooths the observed foul locations by placing a symmetric Gaussian function at each point, estimating the intensity at any location as a weighted average of surrounding fouls. The estimated density function is given by:

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}, \frac{y - y_i}{h}\right) \quad (10)$$

where:

- n is the number of fouls observed,
- (xi, yi) are the coordinates of the i-th foul,
- h is the bandwidth parameter controlling the degree of smoothing.

The bandwidth h determines the scale at which spatial structure is captured: smaller values emphasize local variations, while larger values produce smoother surfaces. The value of h was selected via grid search to minimize model deterioration while preserving visual interpretability.

A preliminary quadrat test was performed to confirm that the distribution of fouls across the pitch was not due to CSR. By dividing the pitch into a grid of quadrats and counting the number of fouls in each, the variance-to-mean ratio of the counts are calculated. A ratio significantly different from one would reject the null hypothesis of CSR, providing the necessary statistical evidence to justify the use of a non-homogeneous model.

This non-parametric method allows flexible estimation of foul densities across the pitch, enabling the detection of spatial patterns and clustering tendencies.

To provide a visual and comparative context for the estimated foul patterns, the thinning algorithm was employed to simulate a single, representative game for each of the five leagues. This simulation process is a key extension of our NHPPP modelling, turning a descriptive model into a generative one.

The steps are as follows:

1. Calculate intensity.
2. Apply the thinning algorithm to generate the points. Define a constant, uniform intensity ($\lambda\, max$) that is greater than or equal to the maximum value of our scaled intensity function.
3. Simulate a homogeneous Poisson process with this uniform intensity across the entire pitch. This gives us an initial set of "candidate" points. For each candidate point, random number between 0 and 1 is generated. If this random number is less than or equal to the ratio of the scaled intensity at that point to the maximum intensity, $\frac{\lambda(x,y)}{\lambda} max$ it is kept. All other points are "thinned" or rejected.

The final set of "kept" points represents a simulated foul pattern for a single, typical game in that league. This provides a clear, visual representation of the combined effect of each league's foul frequency and spatial distribution, serving as a powerful tool for qualitative comparison and insight.

Heatmaps have then been produced by looking at the ratio difference in intensity between countries:

$$R(x,y) = \frac{\lambda_2(x,y)}{\lambda_1(x,y) + \varepsilon} \tag{11}$$

where:

- $\lambda_1(x,y)$ and $\lambda_2(x,y)$ are the spatial intensity functions at location (x,y) for two different groups or conditions,
- $\varepsilon > 0$ is a small constant added to the denominator to prevent division by zero or numerical instability during calculation.

This ratio $R(x,y)$ quantifies the relative intensity of events at each spatial location, allowing comparison between the two intensity surfaces while ensuring computational robustness.

E. Zone Identification Using Gaussian Mixture Models (GMMs)

To identify natural clusters of foul activity across the pitch, GMMs were used. A GMM models the data as a mixture of multiple Gaussian (normal) distributions, providing a flexible, data-driven approach to uncovering spatial groupings in foul locations.

Formally, the probability density of a point $x = (x,y)$under a GMM with K components is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \; \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{12}$$

where:

- π_k is the mixing proportion for component k, with $\sum_{k=1}^{K} \pi_k = 1$
- μ_k is the mean location of component k.
- Σ_k is the covariance matrix of component k, and
- $N(\cdot)$ denotes the bivariate normal distribution.

The optimal number of components K was selected using the BIC, which balances model fit and complexity. The model with the lowest BIC score was chosen. Clustering was performed on the combined dataset across all leagues, ensuring the resulting zones were consistent and suitable for cross-league comparison.

Although GMMs provided a useful exploratory framework, the resulting clusters did not always align with football-specific pitch features (e.g., box, wings, central areas). Therefore, the GMM output was used as a guide for zone placement, but zones were subsequently adjusted to ensure tactical and structural relevance as well as spatial comparability for statistical testing.

GMMs are implemented using the GaussianMixture class from the scikit-learn Python library (Pedregosa et al., 2011), which provides a robust and widely used framework for unsupervised probabilistic modelling.

F. Permutation Testing for Zone Comparison

To compare foul distributions between leagues across predefined pitch zones, permutation testing was employed using the energy test as the test statistic.

The energy distance between two samples $X = \{x1, x2, \dots, xm\}$ and $Y = \{y1, y2, \dots, yn\}$, where each $xi, yj \in Rd$, is defined as:

$$\mathcal{E}(X,Y) = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}|x_i - y_j| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|x_i - x_j| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j| \tag{13}$$

Where $\|\cdot\|$ denotes the Euclidean distance between points.

This statistic measures the distance between the distributions of the two samples and is zero if and only if they have identical distributions.

G. Stouffer's Z Method (Weighted)

To assess whether there was consistent spatial foul behaviour across leagues, Stouffer's Z Method (Weighted) was used to combine the results of the zone-level permutation tests.

Consider L groups, and for each spatial zone z, let the Z-score obtained from the permutation test for group l be $Z_{l,z}$ where $l = 1, 2, \dots, L$. The foul count in zone z for group l is $n_{l,z}$which reflects the stability of the estimate.

The combined Z-score for zone z, denoted Z_z^* is computed as:

$$Z_z^* = \frac{\sum_{l=1}^{L} w_{l,z} Z_{l,z}}{\sqrt{\sum_{l=1}^{L} w_{l,z}^2}} \tag{14}$$

where the weights $w_{l,z}$ are given by:

$$w_{l,z} = \sqrt{n_{l,z}} \tag{15}$$

This weighting assigns greater influence to groups with larger foul counts in each zone, improving the reliability of the combined statistic. The normalization in the denominator ensures Z_z^* follows a standard normal distribution under the null hypothesis, allowing valid statistical interpretation.

IV. Results

A. Data overview

Table 1 displays the average number of fouls committed per match across the five major European leagues in the 2015/2016 season. These values were

calculated by aggregating foul events from the event-level dataset, grouped by league and match, and computing the mean foul count per game. Specifically, for each competition, the total number of fouls was divided by the number of matches played in that league.

This summary provides a baseline comparison of foul frequency across leagues, offering contextual background for the subsequent spatial analysis. Notably, the English Premier League exhibits a substantially lower average foul count (25.03 per game) compared to the other leagues, which all exceed 31 fouls per game. This suggests a more permissive refereeing style in England, consistent with findings reported in prior literature on officiating standards across European competitions.

TABLE 1 - Average fouls per game by league during the 2015/2016 season

Competition	Average Fouls Per Game
Italy - Serie A	33.32
Germany - 1. Bundesliga	32.16
Spain - La Liga	31.94
France - Ligue 1	29.73
England - Premier League	25.03

Poisson Regression Analysis and Multiple Testing Correction

Table 2 reports the results of pairwise Poisson regression tests comparing the average foul counts across the five major European leagues. Each comparison tests the null hypothesis that two leagues exhibit equal mean foul counts per match. P-values were adjusted using the Holm–Bonferroni method at a 5% FWER to control for multiple comparisons.

The analysis reveals statistically significant differences in foul counts between several league pairs. Most notably, the English Premier League differs significantly from all other leagues (p-adjusted = 0 in all comparisons), confirming its lower average foul rate. Spain, and Germany are not statistically different from one another.

These results provide a statistically robust basis for comparing the spatial distribution of fouls, as leagues differ not just in where fouls occur, but also in how frequently they are called.

TABLE 2 - Pairwise comparison of average foul counts per match across leagues using Poisson regression with Holm-Bonferroni correction.

Comparison	Adjusted p-value	Significant
Spain vs England	0.00E+00	TRUE
Spain vs France	1.76E-07	TRUE
Spain vs Italy	2.51E-03	TRUE
Spain vs Germany	6.02E-01	FALSE
England vs France	0.00E+00	TRUE
England vs Italy	0.00E+00	TRUE
England vs Germany	0.00E+00	TRUE
France vs Italy	0.00E+00	TRUE
France vs Germany	5.89E-08	TRUE
Italy vs Germany	1.70E-02	TRUE

An initial test for dispersion indicated potential overdispersion (dispersion ratio = 1.41). However, the negative binomial model performed worse (AIC = 16,142) than the Poisson model (AIC = 12,111), so the Poisson model was retained.

Mirrored zone pairs were compared using a Poisson model to test whether the spatial zones could be reduced. An initial dispersion check confirmed an appropriate distribution (dispersion ratio = 1.28). Several mirrored pairs showed statistically significant differences after Holm correction, indicating that foul distribution is not symmetric across the pitch (full results in Appendix Table A1). Consequently, symmetry-based transformations were not applied in subsequent analyses.

B. Spatial Point Process Analysis

To select an appropriate bandwidth, several values were tested using log-likelihood as the performance metric, Appendix figure A1. While the optimal bandwidth was 1.12, other bandwidths did give a similar performance. Appendix Figure A2. shows how a range of different bandwidths interact with La Liga, lower bandwidths produce high granularity, whilst higher bandwidths over smooth. Appendix Table A2 shows the quadrat test by league, this was run at an 80%,20% test train split. A statistically significant deterioration can be seen when testing k=4.

Therefore, bandwidth of k=3 has been selected as a reasonable balance point. This selection helps allow easier interpretability of charts and graphs throughout the project.

This selected bandwidth of 3 is used throughout the analysis for consistency purposes.

Figure 1. shows the intensity on a one-way basis in La Liga. Charts for the other leagues can be found in the appendix. There is consistent behaviour as might be expected when considering the general flow of football game with a lot of fouls happening the middle of the pitch. The direction of attack is marked, please assume this to be the standard direction in all figures going forward.

There is an observable shift with higher foul intensity being given in the attacking half of the pitch down the right with the opposite happening on the left-hand side of the pitch. This may be linked to players being predominantly right footed. There is also a spike of fouls in the defensive box.

Figure A3. in the appendix shows all the leagues. Broadly speaking the intensity patterns look consistent from this initial viewpoint.

Figure A5. in the appendix shows a distribution of fouls for as simulated match.

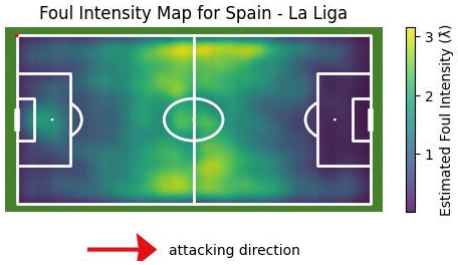

Figure 1. Estimated foul intensity map for La Liga during the 2015/2016 season.

Figure 2. applies the ratio difference between Spain and England. Charts containing the other league comparisons can be found in the appendix. The baseline country is the first named country in the title. Where lighter colours exist the baseline country has relatively higher intensity. Where darker colours exist the baseline country has lower intensity.

England has lower intensity in the defensive end, but a shift can be observed further up the pitch with some noise in the attacking box.

Figure A4. in the appendix shows all the league comparisons. It can be seen here that English refs are much less prone to giving fouls in the attacking box but more prone to giving fouls outside the attacking box. Conversely, Italian refs are much more prone to giving fouls in the attacking box but less so outside the attacking box. French and Italian refs appear to give the most fouls in and around the defensive box.

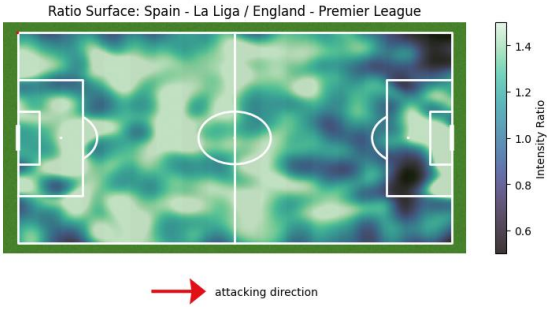

Figure 2. Ratio surface comparison of foul intensity between La Liga and the Premier League across 2015/2016 season.

C. GMM

To define appropriate spatial zones for the Poisson point process analysis, foul data were first aggregated overall. A GMM was applied after KDE, with model fit evaluated via the BIC. The analysis indicated that dividing the pitch into eight zones optimizes model performance (see Appendix Figure A6).

Figure 3. shows how these suggested zones look. It's necessary to adjust the approach here to create appropriate zones, however this a useful guide.

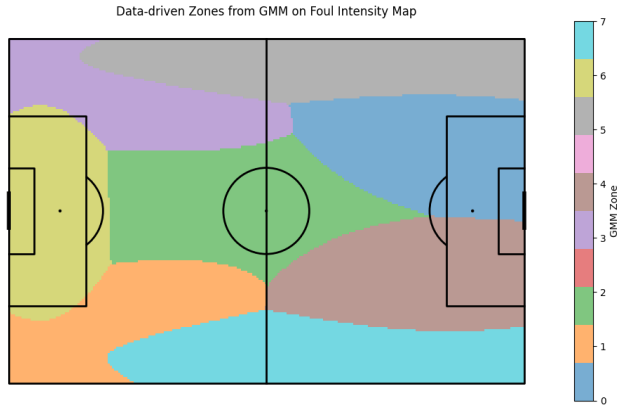

Figure 3. Data-driven zones from GMM clustering of foul intensity.

In figure 3. relatively clear splits can be seen along the horizontal axis of the pitch. Each end of the pitch is roughly split horizontally with the defensive box also being unique area.

Taking figure 3. as a guide, Figure 4. shows the split of zones used using the combined approach.

Due to boxes being unique areas of the pitch, with penalties in the attacking box and goalkeeper handling in the defensive box. Other set pieces also lead to unique situations, it makes sense to test these areas on their own merits.

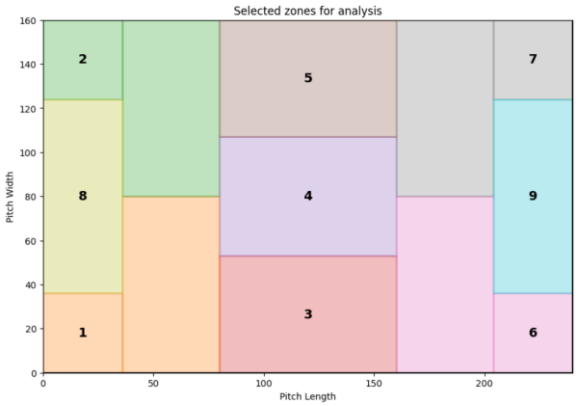

Figure 4. The final zones selected for analysis.

D. Permutation Testing

With these defined zones, I applied permutation testing by comparing different leagues across them.

Figure 5. shows the difference between La Liga and the Premier League. Charts containing the other league comparisons can be found in the appendix. If there is a statistically significant difference the zone shows as white. It can be seen there is a significant difference in the centre of the pitch and the left attacking zone.

Figures A7. and Figure A8. in the appendix show all the league based zonal comparisons. Figure A8. has the true p values for additional context. Germany vs Spain has no statistically significant zones, while Germany vs England and Spain vs England have 2 statistically significant zones. These comparisons have the lowest number of areas with statistical differences.

There were no statistically significant differences observed in either box.

France notably has numerous statistical differences in the defensive half.

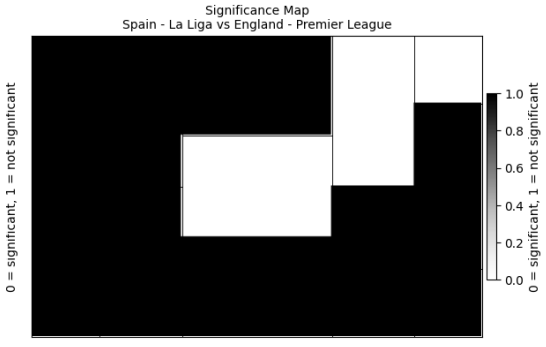

Figure 5. significance map of zonal permutation testing, comparing foul distributions between La Liga and the Premier League.

E. Stouffer's Z Method (Weighted)

To assess whether consistent spatial differences existed between leagues, Stouffer's Z method was used to aggregate the results of the zone-level permutation tests.

This approach combined the p-values from each individual zone comparison into a single, overall measure of significance for each league pair. A weighted version of the method was implemented, assigning greater influence to zones with higher foul counts to improve the reliability of the combined statistic. The results of this analysis are summarized in Table 3.

The results revealed a clear pattern of spatial divergence, with France and Italy showing a statistically significant spatial difference when compared to all other leagues.

However, not all league pairings showed significant overall differences. The comparisons of Germany vs. England and Germany vs. Spain, indicated no statistically significant difference between these leagues. These findings, as shown in Table 3, suggest that while these leagues may differ in their total foul counts, the underlying spatial distributions of fouls are more similar. This highlights a convergence in spatial behaviour that is not captured by simple aggregate foul counts.

TABLE 3 - Stouffer Z method, showing overall spatial difference between league pairs.

Stouffer Z method, showing overall spatial difference between league pairs		
comparison	**Stouffer p corrected**	**significant**
England vs Italy	1.80E-04	TRUE
France vs England	1.21E-02	TRUE
France vs Italy	3.41E-07	TRUE
Germany vs England	8.90E-02	FALSE
Germany vs France	1.14E-05	TRUE
Germany vs Italy	2.10E-02	TRUE
Germany vs Spain	5.41E-01	FALSE
Spain vs England	2.65E-02	TRUE
Spain vs France	1.17E-06	TRUE
Spain vs Italy	2.39E-02	TRUE

V. Conclusions

This study confirms significant quantitative and spatial differences in foul behaviour across the top five European football leagues during the 2015/16 season. The analysis revealed that foul locations are not uniformly distributed and that while the English Premier League exhibits a substantially lower foul count, it's spatial patterns potentially exhibit some similarities compared to Germany. This shows the benefit in comparing surface level vs granular data. This work's value lies in its application of spatial point process models, GMMs, and robust inference techniques to analyse football events, providing a foundation for future research that could incorporate temporal data and referee-level identifiers to further disentangle the observed effects.

A. limitations

The study is limited as the analysis covers only one season, which restricts the ability to account for temporal changes in refereeing style, tactical evolution, or regulatory interventions (e.g., the later introduction of video assistant referees (VAR)). As the data are from 2015/2016 conclusions aren't immediately applicable.

B. Practical implication

These results carry important implications for multiple stakeholders in football. Coaches and analysts could exploit the identified spatial foul tendencies when preparing opposition strategies—for example, targeting areas where referees in certain leagues are less likely to call fouls in the attacking third of the pitch. League bodies and refereeing committees might use these findings to benchmark and monitor consistency in officiating standards across competitions. For clubs, understanding league-specific foul profiles could influence recruitment. Beyond the pitch, broadcasters and pundits could use spatial foul maps to add richer tactical commentary, while gambling and predictive modelling firms might incorporate foul distributions into pricing models for bookings, free kicks, or even match outcome probabilities.

VI. Future Work

This study provides a foundation for spatially analysing foul behaviour across football leagues, but several extensions could enhance its scope and depth:

A. Multi-Season and Longitudinal Analysis

Expanding the dataset to include multiple seasons would allow for temporal modelling of foul patterns, capturing trends in officiating style, tactical evolution, or the impact of regulatory changes (e.g., VAR implementation). Obtaining more recent data would allow more relevant analysis.

B. Inclusion of Referee-Level Data

As the current dataset lacks multiple seasons, this limits the ability to account for individual officiating tendencies as sample sizes are very small. Incorporating referee metadata could enable hierarchical modelling or fixed-effects approaches to isolate league-level effects from referee-specific behaviour.

C. Player and Team-Level Covariates

Future models could incorporate player positions, formations, and match context (e.g., game state, possession sequences) to better control for tactical influences on foul locations and frequencies.

D. Foul Type

Understanding how foul type (e.g. standing tackle, slide tackle) interacts with foul rates could bring improved behavioural understanding.

E. Advanced modelling approaches

Integratring techniques such as Bayesian hierarchical frameworks could quantify uncertainty more robustly and support adding more complex functionality to the frame work.

VII. References

A. Authors and Affiliations

[1] Akaike, H. (1974) 'A new look at the statistical model identification', IEEE Transactions on Automatic Control, 19(6), pp. 716-723.

[2] Aslan, B. and Zech, G. (2005) 'Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding', Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 537(3), pp. 626–636.

[3] Baddeley, A., Turner, R., Møller, J. and Hazelton, M. (2005) 'Residual analysis for spatial point processes', Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(5), pp. 617–666.

[4] Bialkowski, A., Lucey, P., Carr, P., Yue, Y. and Matthews, I. (2014) 'Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviours', in: Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference, Boston, MA, USA.

[5] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society: Series B (Methodological), 39(1), pp. 1–22.

[6] Harris, C.R. et al. (2020) 'Array programming with NumPy', Nature, 585, pp. 357–362.

[7] Hilbe, J.M. (2011) Negative Binomial Regression: Overdispersion. 2nd edn. Cambridge: Cambridge University Press.

[8] Holm, S. (1979) 'A simple sequentially rejective multiple test procedure', Scandinavian Journal of Statistics, 6(2), pp. 65–70.

[9] Hunter, J.D. (2007) 'Matplotlib: A 2D graphics environment', Computing in Science & Engineering, 9(3), pp. 90–95.

[10] Lewis, P.A.W. and Shedler, G.S. (1979) 'Simulation of nonhomogeneous Poisson processes by thinning', Naval Research Logistics Quarterly, 26(3), pp. 403-413.

[11] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd edn. London: Chapman and Hall/CRC.

[12] McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', in: Proceedings of the 9th Python in Science Conference, pp. 51–56.

[13] Nazarudin, N.A., Osman, N., Yusof, N.A. and Rahman, R.A. (2024) 'Disciplinary measures defining referee activity in top-European football leagues: A cross-sectional investigation', Journal of Physical Education and Sport, 24(1), pp. 73–81.

[14] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. and Giannotti, F. (2019) 'A public data set of spatio-temporal match events in soccer competitions', Scientific Data, 6(1), p. 236.

[15] Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830.

[16] Rezek, I. and Roberts, S. (2005) 'Ensemble hidden Markov models with extended observation densities for biosignal analysis', *Neural Computation*, 17(10), pp. 2232–2255.

[17] Ripley, B.D. (1977) 'Modeling Spatial Patterns', Journal of the Royal Statistical Society: Series B (Methodological), 39(2), pp. 172-212.

[18] Sapp, R.M., Spangenburg, E.E. and Hagberg, J.M. (2018) 'Trends in aggressive play and refereeing among the top five European soccer leagues', Journal of Sports Sciences, 36(12), pp. 1346–1354.

[19] Schwarz, G. (1978) 'Estimating the dimension of a model', Annals of Statistics, 6(2), pp. 461–464.

[20] Seabold, S. and Perktold, J. (2010) 'Statsmodels: Econometric and Statistical Modelling with Python', in: Proceedings of the 9th Python in Science Conference, pp. 57-61.

[21] StatsBomb (2018) StatsBomb Open Data [data set]. StatsBomb Services Ltd. Available at: https://github.com/statsbomb/open-data [Accessed 20 August 2025].

[22] Székely, G.J. and Rizzo, M.L. (2004) 'Testing for equal distributions in high dimension', InterStat, 5(16), pp. 1–6.

[23] Székely, G.J. and Rizzo, M.L. (2013) 'Energy statistics: A class of statistics based on distances', Journal of Statistical Planning and Inference, 143(8), pp. 1249–1272.

[24] Virtanen, P. et al. (2020) 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', Nature Methods, 17(3), pp. 261–272.

[25] Wassenaar, J. (2021) mplsoccer: A Python library for plotting soccer/football charts [Computer software]. Available at: https://mplsoccer.readthedocs.io/en/latest/ [Accessed 20 August 2025].

[26] Whitlock, M.C. (2005) 'Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach', Journal of Evolutionary Biology, 18(5), pp. 1368–1373.

Appendix

TABLE A1 - Pairwise comparison of average foul counts per mirrored zone pair using Poisson regression with Holm-Bonferroni correction

Zone Pair	Adjusted p-value	Symmetric?
A_pair	0.026	No
B_pair	0.104	Yes
C_pair	0.030	No
D_pair	0.328	Yes
E_pair	0.197	Yes

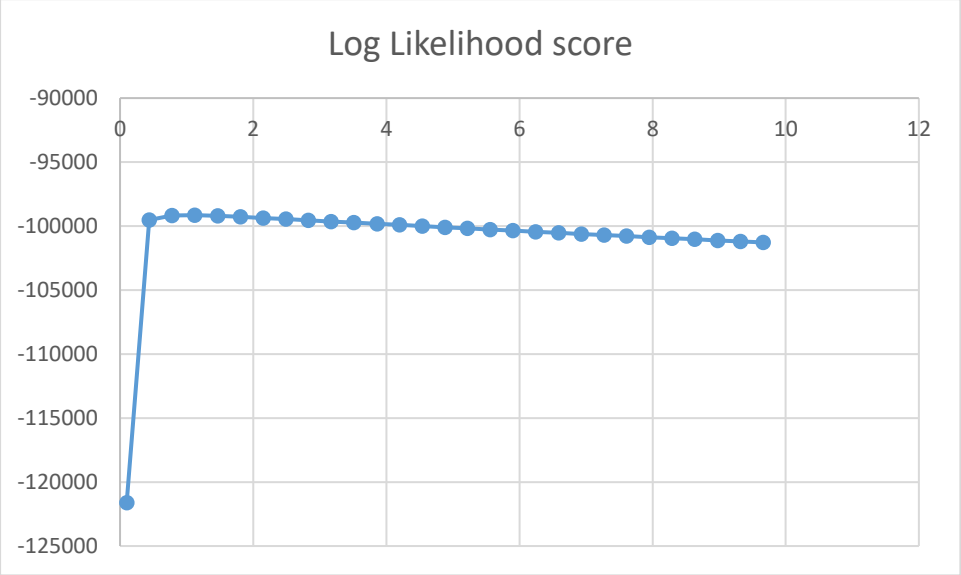

Figure A1: Estimated foul intensity by league

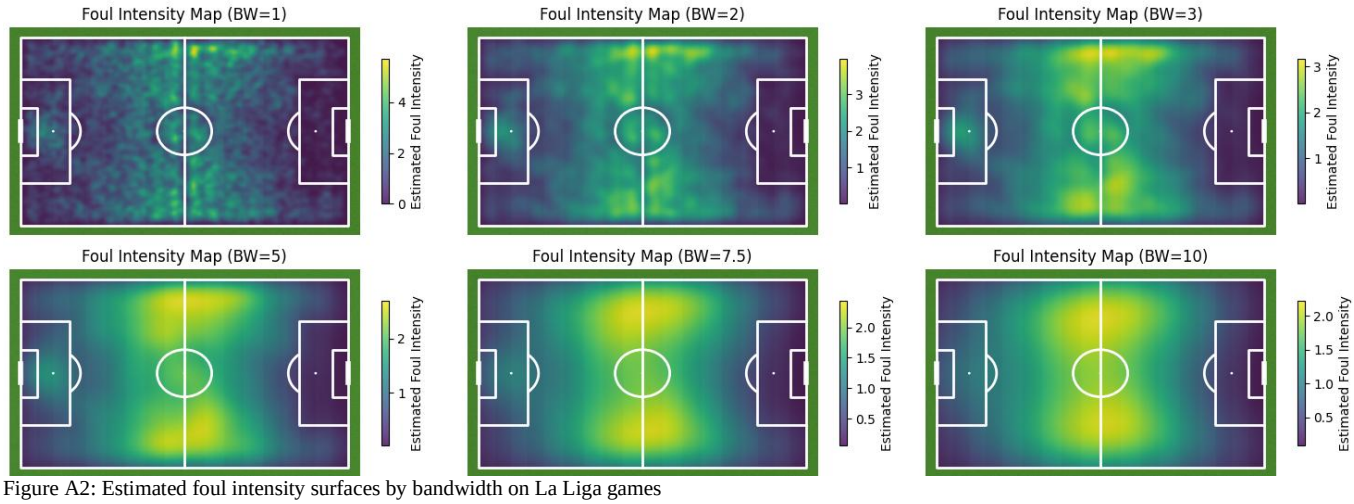

Figure A2: Estimated foul intensity surfaces by bandwidth on La Liga games

TABLE A2 - Quadrat test by bandwidth across leagues

competition	bandwidth	p_value	Significant
Germany	1	0.3431	FALSE
Germany	2	0.4005	FALSE
Germany	3	0.3803	FALSE
Germany	4	0.2692	FALSE
Germany	5	0.1120	FALSE
Spain	1	0.7459	FALSE
Spain	2	0.7655	FALSE
Spain	3	0.6721	FALSE
Spain	4	0.4187	FALSE
Spain	5	0.1177	FALSE
France	1	0.8622	FALSE
France	2	0.9189	FALSE
France	3	0.9369	FALSE
France	4	0.8844	FALSE
France	5	0.6385	FALSE
England	1	0.1218	FALSE
England	2	0.1121	FALSE
England	3	0.0711	FALSE
England	4	0.0254	TRUE
England	5	0.0039	TRUE
Italy	1	0.2204	FALSE
Italy	2	0.1800	FALSE
Italy	3	0.1135	FALSE
Italy	4	0.0438	TRUE
Italy	5	0.0080	TRUE

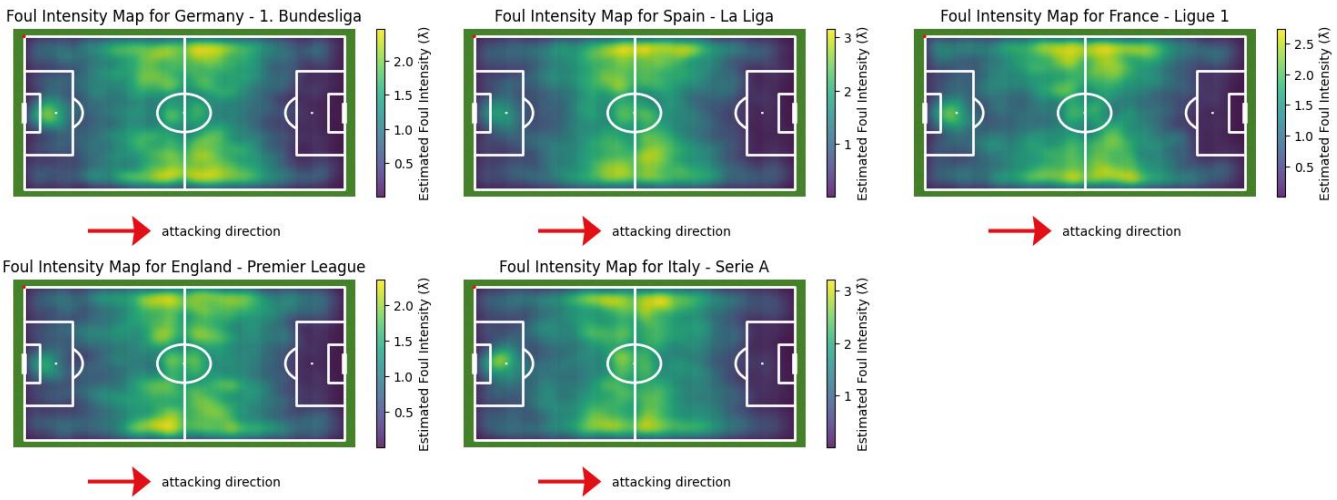

Figure A3: Estimated foul intensity surfaces by league

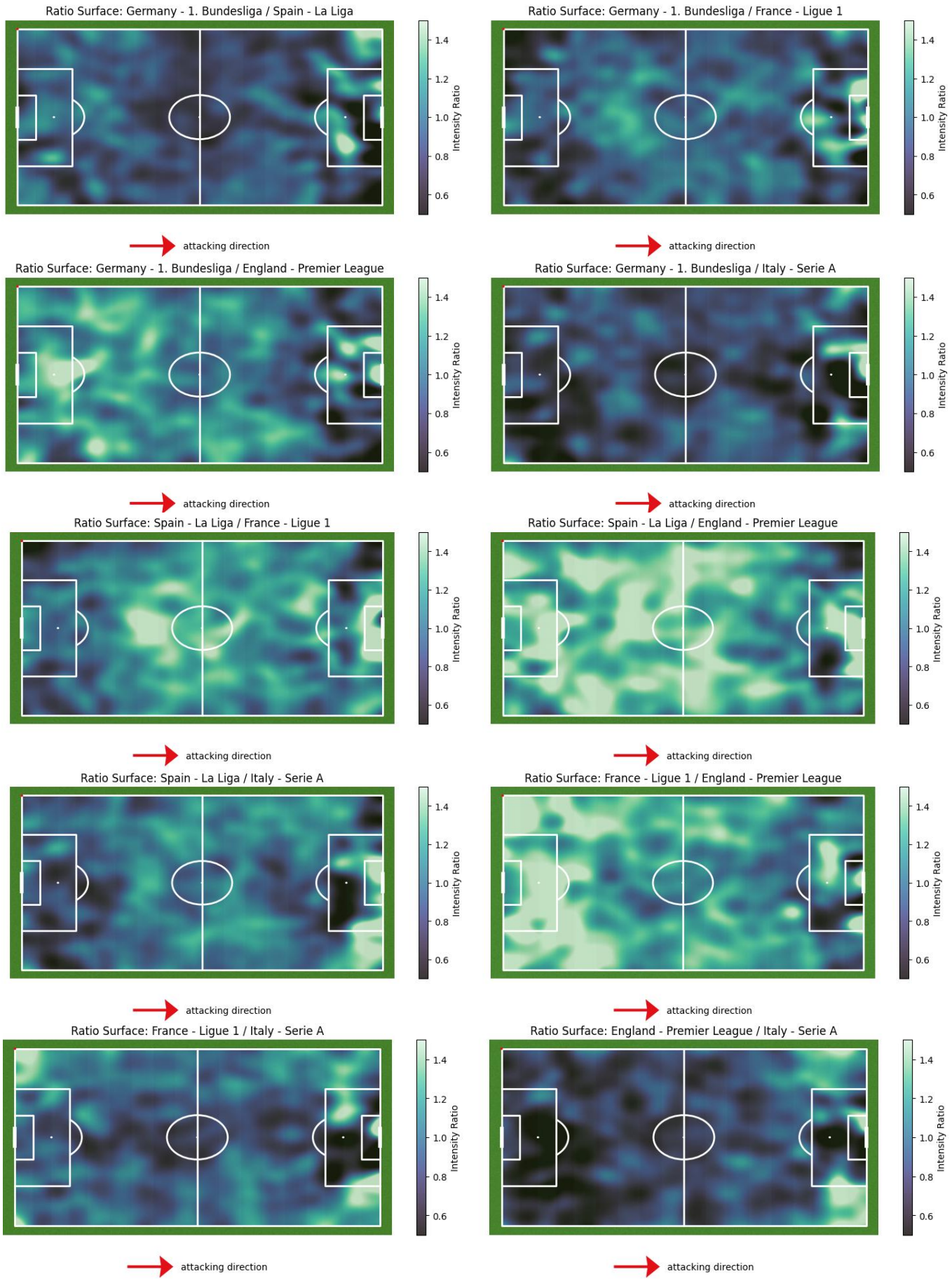

Figure A4: Estimated foul intensity ratio surfaces comparison by league.

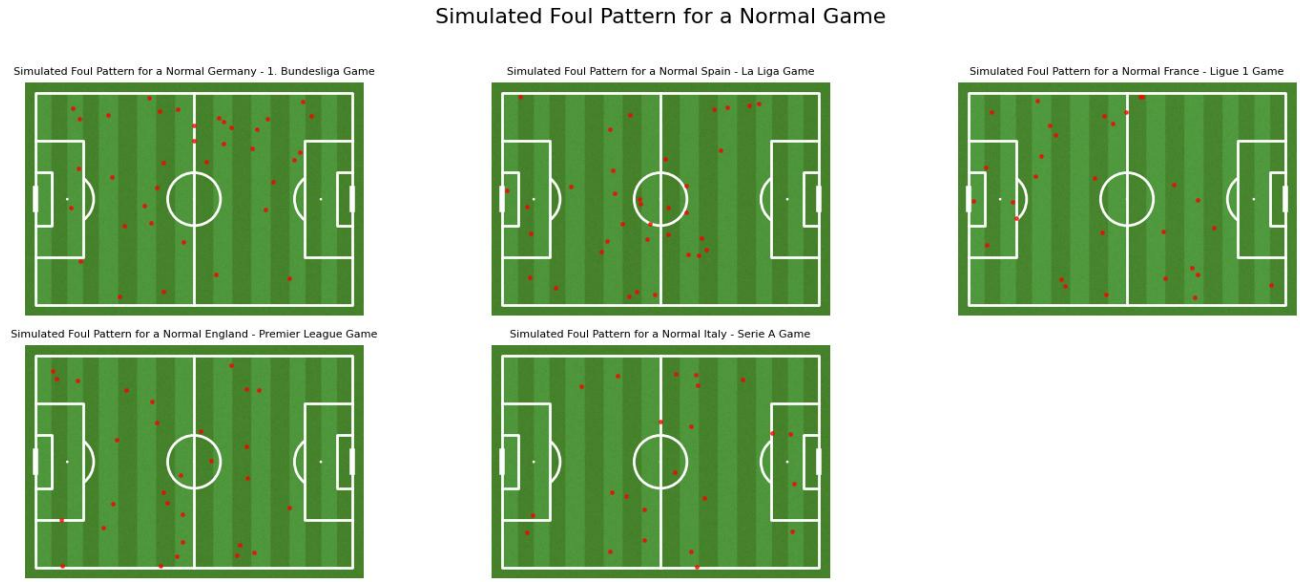

Figure A5. Simulated foul distribution for a game in each league

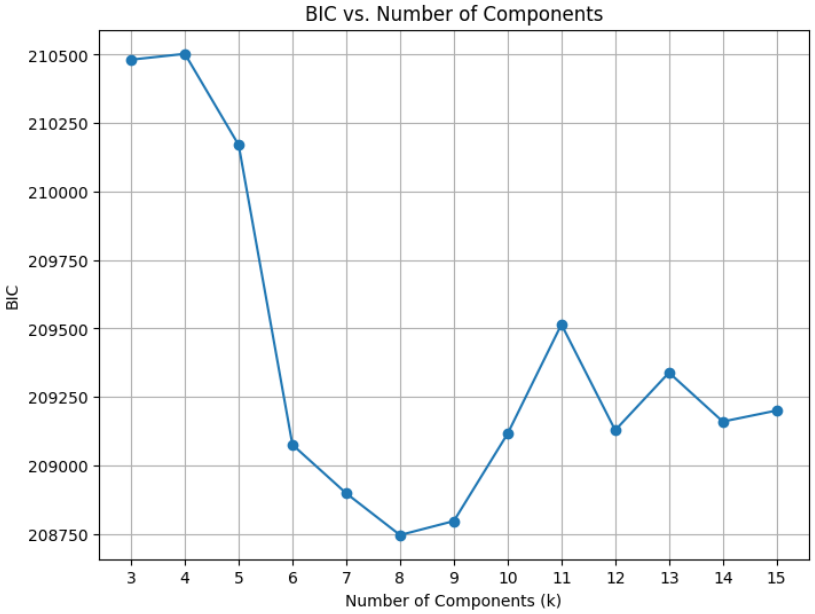

Figure A6. BIC values for GMM with varying numbers of components.

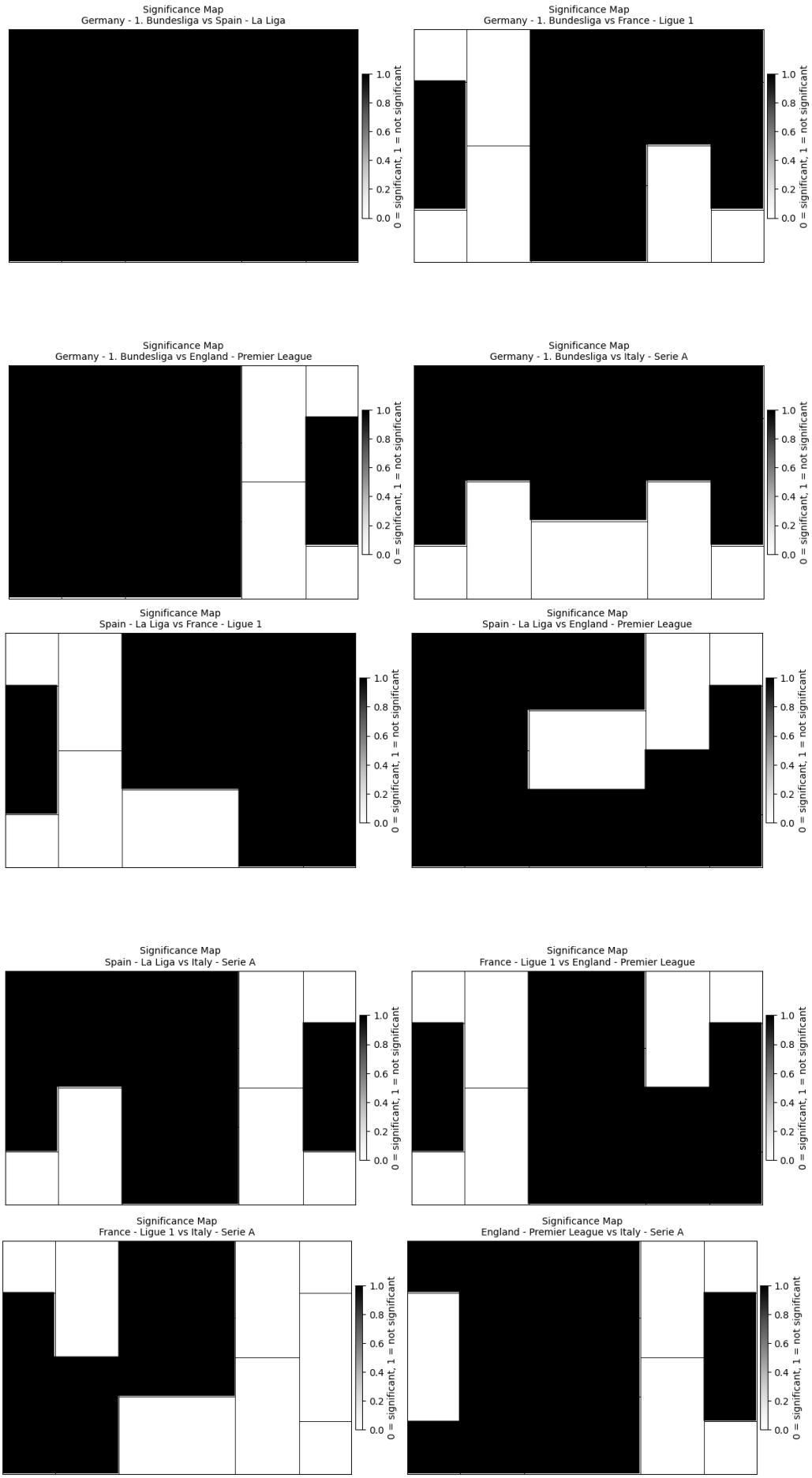

Figure A7: Zonal significance maps comparing foul distributions between league pairs.

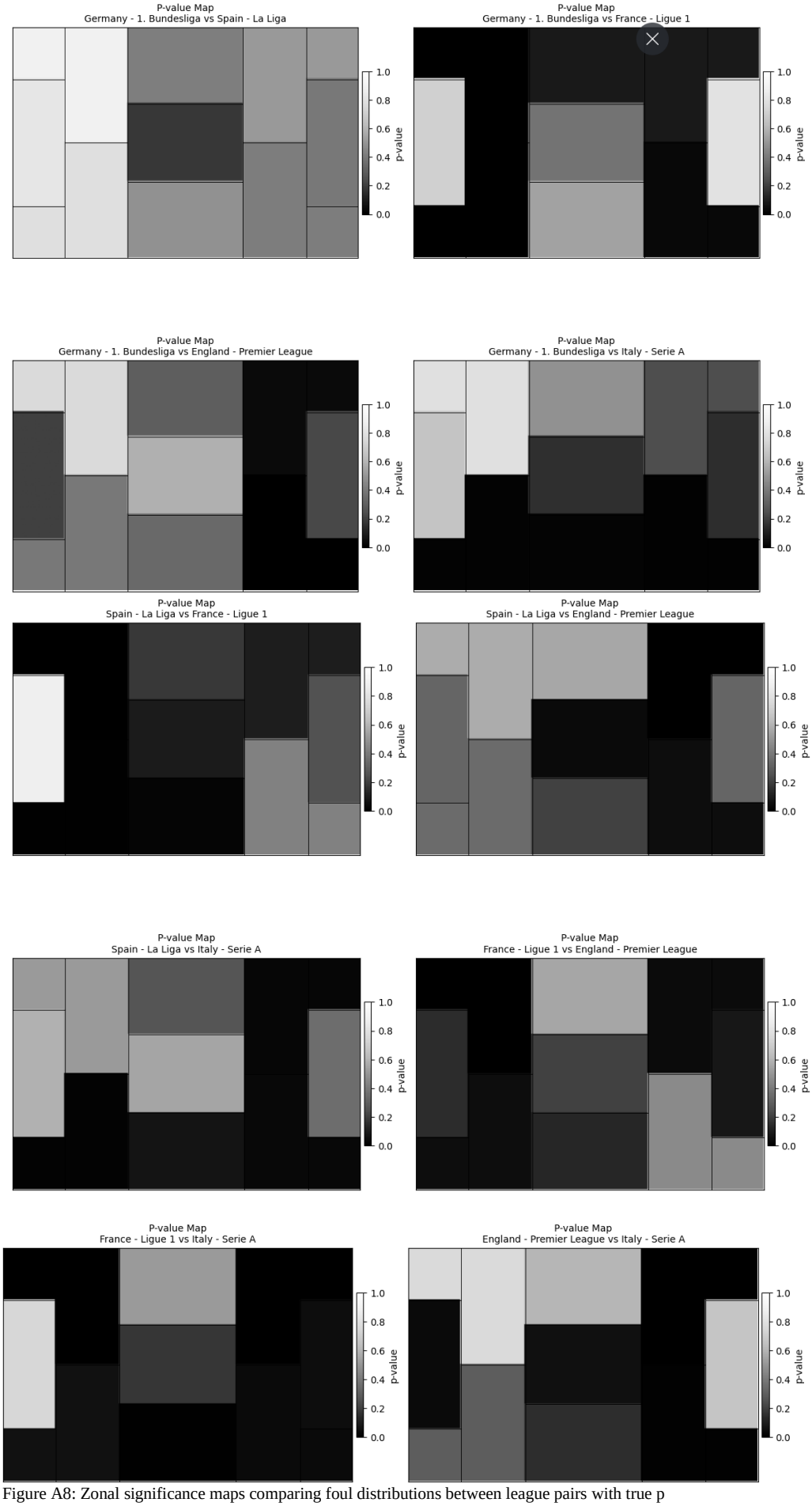

Figure A8: Zonal significance maps comparing foul distributions between league pairs with true p