Spatial Analysis of Soccer Match Events Across European Leaguess

Undi Trivedh Venkata Sai (IMT2023002)
AID 843 STDA - I
IIIT Bangalore, India
Trivedh.Undi@iiitb.ac.in

R Sreenivasa Raju(IMT2023122)
AID 843 STDA - I
IIIT Bangalore, India
Sreenivasa.Raju@iiitb.ac.in

Rajdeep Alapati (IMT2023592)
AID 843 STDA - I
IIIT Bangalore, India
Alapati.Rajdeep@iiitb.ac.in

***Abstract*—This paper presents an exploratory spatial data analysis of a large-scale soccer match event dataset spanning five top tier European leagues England, Spain, Italy, Germany, and France sourced from a publicly available Kaggle repository. The dataset encodes over 643,000 on pitch events per league with precise spatial coordinates normalised to a standardised pitch coordinate system. Shot events are selected as the primary response variable due to their tactical significance and spatial concentration. The pitch is discretised into a regular 11×11 grid and shot density per cell is used as the areal unit for spatial statistical analysis. Three global measures Global Moran's I, Geary's C, and Getis-Ord G are computed for each league to quantify spatial autocorrelation, and Local Moran's I (LISA) along with Getis-Ord G^* are used to identify statistically significant local clusters and hotspots. Spatial heterogeneity is examined through coefficient of variation analysis and Levene's and ANOVA tests across pitch thirds. Geographically Weighted Regression (GWR) is applied to the Spain dataset as a case study. Three regression frameworks OLS, Spatial Lag (GM_Lag), and Spatial Error (GM_Error) are evaluated on both England and the cross league datasets with 80/20 holdout validation. Italy and Spain exhibit the strongest positive spatial autocorrelation in shot placement (Moran's I = 0.391 and 0.325 respectively), while France and Germany show weaker but still significant clustering. GWR applied to Spain achieves $R^2 = 0.414$, substantially outperforming OLS ($R^2 = 0.114$) and confirming spatial non-stationarity in the shot generation process. The results demonstrate that spatial statistical methods provide a principled framework for understanding tactical structure in professional soccer.**

***Index Terms*—Spatial Data Analytics, Soccer Events, Moran's I, LISA, Getis-Ord G^*, Spatial Regression, GWR, Shot Density, Tactical Analysis**

I. Dataset Description

A. Overview

The Soccer Match Event Dataset [1] is a publicly available collection on Kaggle containing match event records from five top-tier European football leagues: England (Premier League), Spain (La Liga), Italy (Serie A), Germany (Bundesliga), and France (Ligue 1). The data originates from Wyscout event feeds and captures every discrete on-pitch action like passes, shots, duels, fouls, and so on recorded during competitive league matches over a full season.

The England dataset, used for in-depth single-league analysis, contains 643,090 event records after cleaning. The four remaining leagues each contain between approximately 550,000 and 640,000 records. Each record represents a single in-game action associated with a precise location on the pitch. Spatial coordinates are encoded as percentages of pitch length (`pos_orig_x`) and pitch width (`pos_orig_y`), both ranging from 0 to 100, effectively normalising all matches to a unified $[0, 100] \times [0, 100]$ coordinate frame. This normalisation is the dataset's most important property for spatial analytics: it guarantees that event locations are directly comparable across matches, teams, and leagues, enabling the application of spatial statistics on a meaningful common domain.

From a spatial analytics perspective the dataset constitutes a *marked spatial point pattern*, where event coordinates form the spatial support and the event type (e.g., Shot, Pass, Duel) acts as the categorical mark. When aggregated to a regular grid the data transitions naturally to an *areal lattice* representation suitable for Moran's I, LISA, and regression-based spatial models.

B. Dataset Structure

Each record in the dataset contains the following key attributes. Spatial attributes include `pos_orig_x` and `pos_orig_y`, the normalised pitch coordinates. Event attributes include `eventName`, a categorical label identifying the type of action, and `subEventName`, a finer-grained subcategory. Match and team identifiers (`matchId`, `teamId`) enable team-level and match-level filtering. Additional fields include `matchPeriod` (first or second half), `eventSec` (time of the event), and a `tags` list encoding binary outcome flags such as whether the event was accurate, a goal, or a key pass.

Shot events were selected as the primary unit of analysis for all spatial statistical procedures. Shots are the most tactically consequential event type; they are spatially concentrated near the penalty area and are directly linked to goal-scoring outcomes, making them ideal for detecting spatial structure. Table I presents the total shot counts per league and the associated grid-level density statistics.

C. Why This Dataset Is Suitable for Spatial Analysis

Several properties of the dataset make it well-suited to spatial statistical analysis. First, the coordinate normalisation ensures a common spatial reference frame across all matches and leagues, enabling inter-league comparison without distortion from ground dimensions. Second, with tens of thousands of shot records per league, the dataset provides ample statistical

TABLE I
SHOT EVENT STATISTICS BY LEAGUE

League	Shots	Mean Density	Variance	CV
England	8,450	84.50	2079.57	0.540
Spain	7,977	65.93	1752.18	0.635
Italy	8,806	72.78	1922.54	0.602
Germany	6,896	56.99	1672.12	0.718
France	8,326	68.81	2656.65	0.749

power to detect spatial autocorrelation at the grid-cell level. Third, the spatial concentration of shots near the penalty area introduces a clear anisotropic spatial structure that global and local autocorrelation methods are designed to quantify. Fourth, the presence of multiple event types enables stratified spatial analysis for passing, duels, and shots each exhibit distinct spatial regimes on the pitch, allowing the investigation of spatial stationarity across event categories. Fifth, the multi-league structure supports comparative spatial analysis across different tactical environments.

II. APPLICABLE SPATIAL STATISTICAL ANALYSIS TECHNIQUES

A. Spatial Representation and Grid Construction

Raw point event data were converted to an areal lattice by binning the pitch into a regular grid. The England single-league analysis used a 10×10 grid (100 cells, each spanning 10 pitch units), while the cross-league analysis used an 11×11 grid (121 cells, spanning approximately 9.1 pitch units each). Each cell's shot count was used as the response variable. Grid centroids served as observation locations for spatial weight construction.

Alongside grid aggregation, continuous pitch-level visualisations were generated using Kernel Density Estimation (KDE) and hexbin aggregation for exploratory insight into the raw point pattern structure.

B. Spatial Weights Matrix

Neighbourhood relationships among grid cells were encoded using a k-nearest neighbour (KNN) spatial weights matrix with $k = 4$ (each cell connected to its four nearest grid neighbours). The matrix was row-standardised so that each weight satisfies:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_j w_{ij}} \tag{1}$$

For the England point-level Moran's I computation on event type labels, a KNN matrix with $k = 8$ was constructed from a 5,000-point subsample. A distance-band weights matrix (threshold = 15 pitch units) was additionally tested in the cross-league analysis to assess robustness of autocorrelation estimates.

C. Global Spatial Autocorrelation

Three complementary global measures were computed to characterise the overall spatial structure of shot density.

Global Moran's I is the standard test for spatial autocorrelation:

$$I = \frac{n}{S_0} \cdot \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{2}$$

where n is the number of grid cells, y_i is the shot density at cell i, $\bar{y}$ is the global mean, w_{ij} are spatial weights, and $S_0 = \sum_i \sum_j w_{ij}$. Values significantly above $E[I] = -1/(n-1) \approx 0$ indicate positive spatial autocorrelation (clustering); values below indicate dispersion.

Geary's C provides a complementary local-contrast measure:

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2S_0 \sum_i (y_i - \bar{y})^2} \tag{3}$$

Values of $C < 1$ indicate positive autocorrelation, $C > 1$ indicates negative autocorrelation, and $C = 1$ corresponds to spatial randomness. Geary's C is more sensitive to local spatial differences than Moran's I.

Getis-Ord Global G tests for the global concentration of high or low values:

$$G(d) = \frac{\sum_i \sum_j w_{ij}(d) y_i y_j}{\sum_i \sum_j y_i y_j}, \quad j \neq i \tag{4}$$

Significance is assessed via Monte Carlo permutation with 999 iterations for all three measures.

D. Local Spatial Autocorrelation: LISA and Getis-Ord G^*

Local Moran's I decomposes global autocorrelation into cell-specific contributions:

$$I_i = z_i \sum_j w_{ij} z_j \tag{5}$$

where $z_i = (y_i - \bar{y})/\sigma$. Cells are classified at $p < 0.05$ into:
High-High (HH) - a dense shooting zone surrounded by other dense zones
Low-Low (LL) - sparse zones surrounded by sparse zones
High-Low (HL) - a local hotspot within a sparse area or Low-High (LH) - a sparse cell surrounded by dense neighbours.

Getis-Ord G^* identifies hotspot and coldspot regions based on the local concentration of values:

$$G_i^{*} = \frac{\sum_j w_{ij} y_j - \bar{y} \sum_j w_{ij}}{\sigma \sqrt{\frac{n \sum_j w_{ij}^2 - (\sum_j w_{ij})^2}{n-1}}} \tag{6}$$

Positive significant G^* z-scores indicate hotspots; negative significant z-scores indicate coldspots. LISA and G^* are complementary: LISA identifies spatial outliers as well as clusters, while G^* focuses purely on concentration intensity.

E. Spatial Heterogeneity and Stationarity

Spatial heterogeneity variation in the statistical properties of shot density across the pitch was assessed through: (i) the coefficient of variation (CV = std/mean) computed globally and by pitch third; (ii) Levene's test for homogeneity of variance between pitch halves and thirds; (iii) one-way ANOVA to test for significant differences in mean shot density across

defensive, midfield, and attacking pitch thirds; and (iv) the distribution of local GWR coefficient estimates, which directly quantify spatial non-stationarity in the regression relationship.

F. Spatial Regression Models

Three regression frameworks were implemented and compared.

OLS serves as the non-spatial baseline: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where independence of residuals is assumed. Residual Moran's I is computed post-estimation to diagnose any remaining spatial structure.

Spatial Lag (GM_Lag) incorporates a spatially lagged dependent variable via the Generalised Method of Moments estimator : $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where ρ is the spatial autoregressive coefficient. The GM estimator avoids the $O(n^3)$ matrix inversion required by Maximum Likelihood estimation.

Spatial Error (GM_Error) models spatially correlated unobserved covariates: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}$, where $\boldsymbol{u} = \lambda\mathbf{W}\boldsymbol{u} + \boldsymbol{\varepsilon}$. A significant λ indicates that omitted variables with spatial structure are biasing OLS estimates.

Geographically Weighted Regression (GWR) allows coefficients to vary spatially across the grid: $y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i$, where (u_i, v_i) are the centroid coordinates of cell i and $\beta_k(u_i, v_i)$ is the locally estimated coefficient. An adaptive bisquare kernel was used with bandwidth optimised via AICc minimisation.

III. Experiments

A. Preprocessing

England (single-league analysis). The raw dataset contained 643,090 records. Preprocessing steps were: (1) column selection retaining `matchId`, `teamId`, `eventName`, `pos_orig_x`, and `pos_orig_y`; (2) removal of rows with null coordinates; (3) coordinate clipping to the valid range $[0, 100]$ on both axes to eliminate erroneous out-of-bounds records. The cleaned dataset retained all 643,090 events. Shot events were then isolated (8,450 total) and assigned to a 10×10 grid, yielding 100 non-empty zone cells. A `GeoDataFrame` was constructed using cell centroids as geometry for spatial weight construction.

Cross-league analysis. The Spain dataset (628,550 total events; 7,977 shots) was loaded separately. An 11×11 grid was used across all four leagues (Spain, Italy, Germany, France), yielding 121 grid cells per league. The same coordinate clipping and shot-filtering steps were applied.

B. Exploratory Visualisations

Exploratory analysis of the England dataset produced: a scatter plot of all 643,090 event locations on the standardised pitch (Figure 1), demonstrating uniform pitch coverage; a KDE density surface revealing concentration near the central midfield corridor and both penalty areas; a hexbin aggregation map (Figure 3) highlighting zones of maximum event frequency; zone heatmaps for the full event set and stratified by the top three event types (Pass, Duel, Others on the ball); and team-level KDE contour overlays for the three highest-volume teams, enabling qualitative tactical comparison.

Figure 2 shows the KDE of shot locations for the England dataset, confirming the expected spatial concentration in the attacking third near the penalty area.

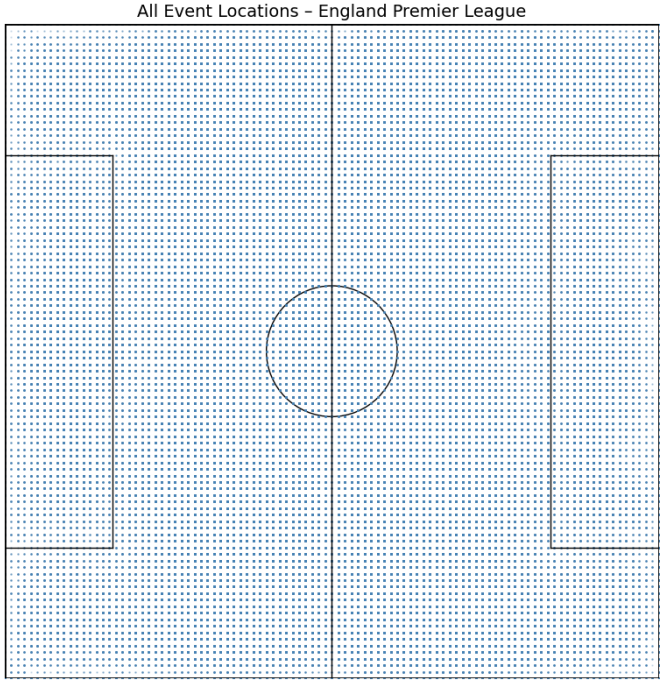

Fig. 1. Scatter plot of all 643,090 match events across the standardised England pitch. Dense coverage across the full $[0, 100] \times [0, 100]$ domain confirms the validity of the normalised coordinate frame for spatial analysis.

C. Spatial Heterogeneity Analysis

For the England shot density grid, descriptive heterogeneity metrics were computed: mean shot density per zone = 84.50, variance = 2079.57, standard deviation = 45.60, and coefficient of variation = 0.540. Levene's test for variance homogeneity between the left and right pitch halves yielded $F = 0.518$, $p = 0.473$, indicating no significant lateral variance difference. The defensive-vs-attacking thirds Levene test yielded $F = 0.261$, $p = 0.611$ also non-significant. These results suggest that while shot density is strongly non-uniform across the pitch (CV = 0.540), the variance structure itself is relatively consistent across spatial partitions in the England dataset.

For the cross-league datasets, ANOVA tests on mean shot density across pitch thirds were significant for Spain ($p < 0.001$), Italy ($p < 0.001$), and Germany ($p = 0.002$), confirming that shot density varies systematically by pitch third in three of the four leagues. France did not show a significant ANOVA result ($p = 0.081$), consistent with its higher CV (0.749) and more spatially dispersed shot distribution. Levene's test was significant only for France ($p = 0.029$), indicating significantly different variances between defensive and attacking thirds in that league.

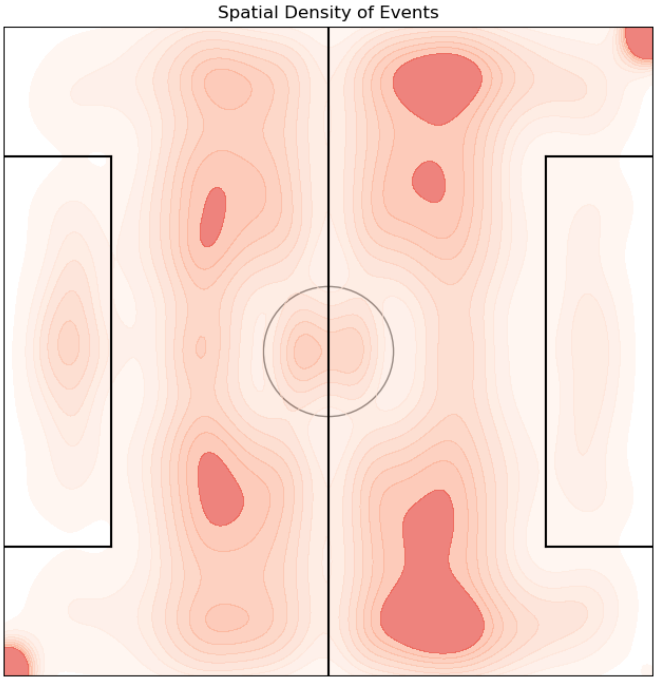

Fig. 2. KDE density surface of events, England Premier League. The dominant concentration in the attacking third (high x, central y) reflects the tactical imperative to shoot from central positions near goal.

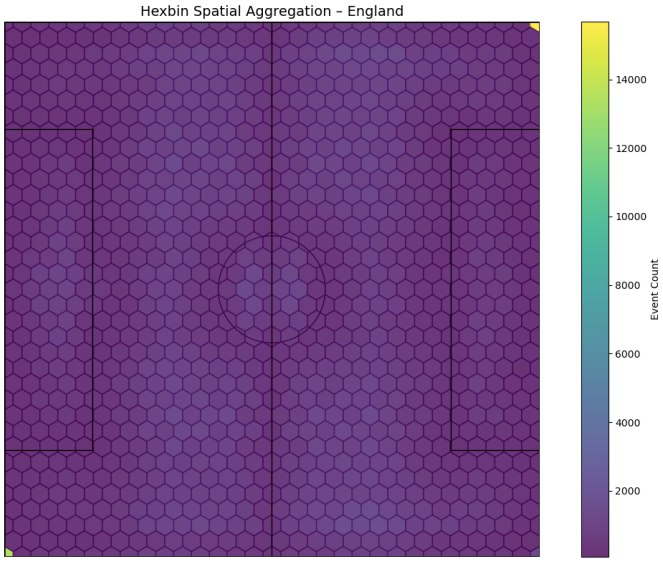

Fig. 3. Hexbin aggregation of all England events.

D. Global and Local Autocorrelation

Global Moran's I, Geary's C, and Getis-Ord G were computed for all four cross-league datasets as well as for England. Moran scatterplots were generated for each league to visualise the spatial lag relationship. LISA cluster maps (Figure 4) were produced for all leagues, and G^* hotspot maps (Figure 5) were generated to show the spatial extent of high-density and low-density zones.

For the England dataset, Global Moran's I was additionally computed stratified by event type (Pass, Duel, Others on the ball) to examine whether different game actions exhibit differing degrees of spatial clustering.

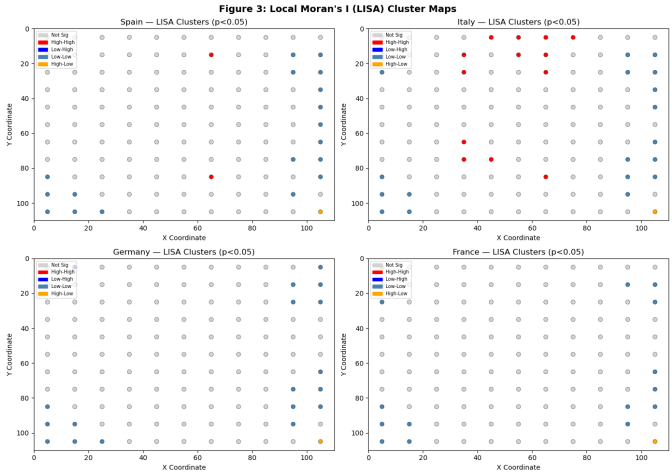

Fig. 4. LISA cluster maps for shot density across all four cross-league datasets. High-High (red) clusters are concentrated in the central attacking third across all leagues; Low-Low (blue) clusters dominate defensive and wide lateral zones.

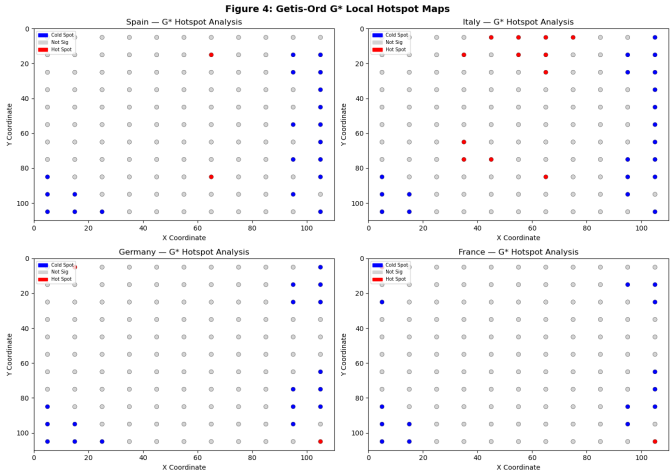

Fig. 5. Getis-Ord G^* hotspot maps for shot density. Positive z-scores (orange/red) identify statistically significant shot hotspots; negative z-scores (blue) identify coldspots. The hotspot pattern is spatially consistent with the LISA High-High clusters.

E. Regression Experiments

All regression models were estimated using an 80/20 train–test split for holdout validation (train $n = 80$, test $n = 20$ for England's 100-cell grid). The dependent variable was shot density per grid cell. Predictors were the cell centroid coordinates x_{coord} and y_{coord}. Spatial weights used in the lag and error models were constructed on the training set only using KNN with $k = 4$, row-standardised.

GWR (Spain case study). GWR was applied to the Spain dataset (121 cells) using standardised centroid coordinates

as predictors. The optimal bandwidth was determined as 51 nearest neighbours via AICc minimisation.

Residual Moran's I was computed on all model residuals to quantify the degree of remaining spatial autocorrelation after model fitting. Residual scatter plots were generated for each model.

IV. Results

A. Global Spatial Autocorrelation

Table II presents the global autocorrelation results for all five leagues. All leagues exhibit statistically significant positive spatial autocorrelation in shot density across all three global measures.

TABLE II
GLOBAL SPATIAL AUTOCORRELATION - SHOT DENSITY BY LEAGUE

League	Moran's I	p	Geary's C	p	Getis G
England	0.1155	0.015	0.7293	0.002	—
Spain	0.3246	0.001	0.5629	0.001	0.0378
Italy	0.3909	0.001	0.5191	0.001	0.0385
Germany	0.1948	0.002	0.6563	0.001	0.0366
France	0.1250	0.025	0.7155	0.001	0.0356

Italy records the highest Moran's I (0.391), indicating the strongest tendency for shots to cluster in contiguous high-density zones. Spain follows closely with $I = 0.325$. Germany ($I = 0.195$) and France ($I = 0.125$) exhibit weaker but still significant clustering. Geary's C values below 1.0 for all leagues independently confirm positive autocorrelation. The Getis-Ord G statistic is uniformly positive and significant, confirming global concentration of high shot counts. The three measures are mutually consistent, strengthening confidence in the finding.

For the England dataset, stratified Moran's I by event type (Table III) reveals that Passes exhibit the strongest clustering ($I = 0.410$, $p = 0.001$), followed by Duels ($I = 0.334$, $p = 0.001$) and Others on the ball ($I = 0.311$, $p = 0.001$). This is noteworthy: passes are more spatially clustered than shots at the zone level, because passing activity concentrates in both midfield build-up zones and final-third attacking corridors, generating two distinct HH poles.

TABLE III
GLOBAL AUTOCORRELATION BY EVENT TYPE - ENGLAND

Event Type	Moran's I	Geary's C	p-value	
			Moran	Geary
Pass	0.4095	0.5165	0.001	0.001
Duel	0.3340	0.5713	0.001	0.001
Others on ball	0.3106	0.5890	0.001	0.001
Shot	0.1155	0.7293	0.015	0.002

B. Local Spatial Autocorrelation

Table IV summarises the LISA cluster distributions across leagues. High-High (HH) clusters represent grid cells with high shot density surrounded by similarly dense neighbours, the tactical hotspots of the game.

TABLE IV
LISA AND GETIS-ORD G^* CLUSTER COUNTS BY LEAGUE

League	HH	LL	HL	LH	G^* Hot	G^* Cold
England	1	15	2	1	—	—
Spain	2	18	1	0	2	20
Italy	13	18	1	0	12	20
Germany	0	17	1	1	2	17
France	0	14	2	0	2	14

Italy stands out with 13 significant HH clusters and 12 G^* hotspots, by far the most concentrated spatial structure of any league. This is consistent with Serie A's historically structured, positional style of play, in which attacks are systematically channelled through the penalty area. Spain shows 2 HH clusters and 2 hotspots. Germany and France show no significant HH clusters under LISA, with only G^* hotspots detected, suggesting weaker local concentration of high-density zones.

Low-Low clusters (LL) are present in all leagues (14–18 cells) and are spatially consistent with deep defensive zones and wide lateral areas where shot events are rare across all matches. The small number of High-Low (HL) spatial outliers (1–2 per league) represent isolated grid cells with relatively high shot counts surrounded by sparse-shooting zones, likely reflecting specific set-piece locations or positional anomalies in the data.

The LISA and G^* maps (Figures 4 and 5) confirm that both methods identify spatially consistent clusters: G^* hotspots align with LISA HH clusters in all leagues where both are detected.

C. Spatial Heterogeneity

Table V presents the spatial heterogeneity statistics across leagues. France exhibits the highest CV (0.749) and is the only league with a significant Levene's test ($p = 0.029$), indicating significantly different variance in shot density across pitch thirds, evidence of stronger spatial non-stationarity in the French league's shot distribution. Germany has the second-highest CV (0.718), consistent with its counter-attacking tactical profile generating shots from diverse pitch positions.

TABLE V
SPATIAL HETEROGENEITY METRICS BY LEAGUE

League	Mean	Variance	CV	Levene p	ANOVA p
England	84.50	2079.57	0.540	0.473	—
Spain	65.93	1752.18	0.635	0.454	< 0.001
Italy	72.78	1922.54	0.602	0.125	< 0.001
Germany	56.99	1672.12	0.718	0.313	0.002
France	68.81	2656.65	0.749	**0.029**	0.081

ANOVA tests across pitch thirds are significant for Spain, Italy, and Germany, confirming that mean shot density varies systematically from the defensive to the attacking third. Specifically, for Spain: defensive third mean = 60.00, midfield = 87.84, attacking third = 48.45, an inverted U-shape pattern consistent with shots concentrating in central midfield-to-penalty-area zones rather than in the deepest attacking cells

(which correspond to extreme tight-angle positions at the byline).

D. GWR Results

GWR was applied to the Spain dataset as a detailed case study of spatial non-stationarity. The global OLS baseline achieved $R^2 = 0.114$. The GWR model improved this to $R^2 = 0.414$ (Adjusted $R^2 = 0.359$), representing a 30-percentage-point improvement that confirms significant spatial non-stationarity in the relationship between pitch location and shot density. The optimal bandwidth was 51 nearest neighbours determined by AICc minimisation.

The locally varying GWR coefficients for x_{coord} (pitch length) and y_{coord} (pitch width) show wide spatial variation across the grid, confirming that the marginal effect of pitch position on shot density is not constant, it differs substantially between the central attacking corridor, the wide lateral zones, and the defensive half. Figure 6 shows the spatial map of GWR local R^2 values.

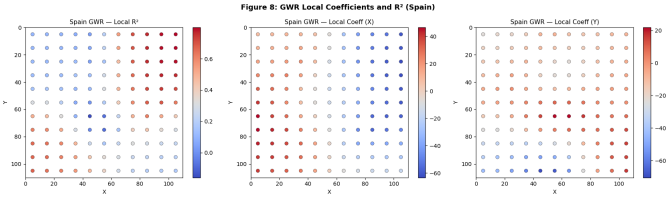

Fig. 6. GWR local R^2 map for Spain. Cells with high local R^2 (warm colours) are primarily in the attacking third, where pitch position is a strong predictor of shot density. Lower R^2 in defensive zones reflects greater ecological heterogeneity in those areas.

E. Regression Model Comparison

Table VI presents the regression model comparison for England. OLS achieves MAE = 30.49 on the test set with $R^2 = 0.076$, indicating that pitch centroid coordinates alone explain limited variance in shot density, consistent with the non-linear, anisotropic nature of shot concentration. The Spatial Lag model's test MAE of 305.70 is an artefact of the GM estimator struggling with the small training sample ($n = 80$): the estimated spatial autoregressive coefficient falls outside the stable $(-1, 1)$ boundary, rendering the spatial lag predictions unreliable. The Spatial Error model (MAE = 31.02) performs comparably to OLS.

TABLE VI
REGRESSION MODEL COMPARISON - ENGLAND (80/20 SPLIT)

Model	Test MAE	R^2	Res. Moran's I	p
OLS	30.49	0.076	0.143	0.010
GM_Lag	305.70	0.047	0.624	0.001
GM_Error	31.02	0.076	0.157	0.008

All three models exhibit statistically significant residual Moran's I ($p \leq 0.01$), confirming that the spatial structure in shot density is not fully captured by pitch centroid coordinates alone. The residual Moran's I of OLS ($I = 0.143$) is lower than that of the failed lag model ($I = 0.624$) but substantially higher than what a well-specified spatial model should produce. This motivates the GWR approach, which by allowing coefficients to vary spatially was able to achieve a substantially higher R^2 in the Spain case study.

Table VII presents the cross-league OLS and spatial error results for comparison.

TABLE VII
CROSS-LEAGUE REGRESSION COMPARISON (OLS AND GM_ERROR)

League	OLS MAE	OLS R^2	Error MAE	Err. Res. I
Spain	31.31	0.114	31.99	0.367
Italy	34.16	0.113	35.29	0.419
Germany	29.95	0.082	30.32	0.254
France	38.48	−0.002	38.63	0.151

V. DISCUSSION

A. Spatial Structure and Tactical Implications

The analysis confirms strong and consistent positive spatial autocorrelation in soccer shot placement across all five European leagues. The HH clusters identified by LISA and G^* are uniformly located in the central attacking third, corresponding to the penalty area and the zones immediately in front of goal. This is the expected tactical outcome: professional teams systematically seek central shooting positions because centrality and proximity to goal are the strongest determinants of shot quality and goal probability. The spatial clustering captured by Moran's I and LISA therefore reflects both the physics of the game and the strategic rationality embedded in professional tactical systems.

The inter-league variation in Moran's I is practically meaningful. Italy ($I = 0.391$) and Spain ($I = 0.325$), leagues historically associated with structured positional play and organised offensive sequences, exhibit the strongest spatial clustering. Germany ($I = 0.195$) and France ($I = 0.125$) leagues with higher game pace, more vertical transitions, and greater tactical variety, show weaker clustering, consistent with shots arising from a more spatially diverse set of positions. England ($I = 0.116$, $p = 0.015$), with the highest-paced league and significant emphasis on wide attacking play, shows the weakest autocorrelation among the five leagues.

B. Did the Chosen Methods Work for This Dataset?

The spatial autocorrelation methods (Moran's I, Geary's C, G/G^*, LISA) worked effectively for this dataset. The normalised coordinate system provided a clean common spatial frame; the high observation counts ensured stable estimators; and the known tactical concentration of shots near the penalty area provided a ground-truth validation against which the cluster maps could be evaluated, the identified HH clusters and G^* hotspots consistently mapped to penalty-area zones, confirming that the methods are detecting genuine tactical structure rather than artefacts.

The GWR model worked well, achieving a substantial improvement in R^2 over OLS in the Spain case study and

identifying interpretable local variation in the shot-location relationship. The standard (non-GW) spatial regression models (GM_Lag, GM_Error) were less successful. The primary limitation was the small number of grid cells (100–121), which is a very small spatial sample for GM estimators. The GM_Lag estimator in particular produced an unstable spatial autoregressive coefficient ($\rho > 1$) in several leagues, rendering lag model predictions unreliable. The GM_Error model performed comparably to OLS in test MAE but did not substantially reduce residual autocorrelation, suggesting model misspecification, the linear relationship between centroid coordinates and shot density is a poor approximation of the true non-linear concentration pattern.

C. Limitations

The primary methodological limitation is the coarseness of the grid representation. A 10×10 or 11×11 grid results in only 100–121 spatial observations, which is a small sample for spatial regression. A finer grid (e.g., $20 \times 20 = 400$ cells) would provide more observations and more stable regression estimates, but at the cost of sparser shot counts in peripheral cells. The regression models also use only pitch centroid coordinates as predictors. Including tactical features such as shot angle from goal, distance to goal, match half, or binary zone indicators for penalty area membership would substantially improve predictive performance and ecological interpretability. Finally, the analysis treats all matches and all time periods equally; incorporating temporal stratification (e.g., first half vs second half, by score state) would reveal whether shot placement changes under different game contexts.

D. What Further Analysis Could Improve Understanding

Several additional analyses not feasible within the scope of this work would deepen understanding of soccer event spatial dynamics.

Distance-based correlograms (Moran's I as a function of distance band) would reveal the effective spatial range of shot clustering analogous to a variogram in geostatistics providing a scale-dependent characterisation of tactical organisation. The Spatial Durbin Model (SDM), which incorporates both the spatially lagged dependent variable and spatially lagged predictors, would offer a richer specification of spatial spillover in shot creation than the standard lag model. Team-stratified LISA maps would allow objective spatial characterisation of team-specific tactical identity for example, distinguishing teams that concentrate shots exclusively through central channels from those using wide attacks. Finally, extending the analysis to a spatio-temporal dimension by incorporating match time (e.g., analysing shot density at 15-minute intervals) would reveal how spatial concentration evolves over the course of a match, potentially capturing tactical shifts in response to scoreline changes.

VI. Conclusion

This study demonstrates the applicability and effectiveness of spatial statistical methods for analysing soccer match event data. By aggregating shot events onto a regular pitch grid and applying a comprehensive suite of global autocorrelation measures (Moran's I, Geary's C, Getis-Ord G), local cluster statistics (LISA, G^*), heterogeneity tests (CV, Levene, ANOVA), and regression models (OLS, GM_Lag, GM_Error, GWR), we establish that shot placement is spatially clustered in all five European leagues analysed. The strength of clustering varies across leagues in a manner consistent with known differences in tactical playing styles: Italy and Spain exhibit the strongest spatial autocorrelation, while England and France show weaker but significant clustering. LISA cluster maps consistently identify the central penalty area as a High-High hotspot, and G^* maps corroborate this finding across all leagues. GWR applied to Spain confirms significant spatial non-stationarity and substantially outperforms OLS ($R^2 = 0.414$ vs 0.114), demonstrating that the relationship between pitch location and shot density is geographically non-uniform. Standard spatial lag and error models were limited by the small number of grid cells; future work with finer grids and richer feature sets would substantially improve predictive performance. Overall, the spatial statistical framework developed here provides a principled, interpretable, and empirically validated approach to extracting tactical structure from large-scale soccer event data.

Team Members and Work Distribution

- **[Rajdeep Alapati] (IMT2023592):** Data preprocessing including column selection, missing value and coordinate bounds cleaning, grid zone assignment, and GeoDataFrame construction. Exploratory visualisations including scatter plots, KDE density maps, hexbin aggregation, zone heatmaps (all events and per event type), and team-level tactical comparison plots. Spatial weights matrix construction and row standardisation.
- **[Undi Trivedh Venkata Sai] (IMT2023002):** Global spatial autocorrelation analysis computation and interpretation of Global Moran's I, Geary's C, and Getis-Ord G for all five leagues. Moran scatterplot generation. Local spatial autocorrelation, Local Moran's I (LISA) cluster computation and cluster map generation, Getis-Ord G^* hotspot and coldspot mapping. Stratified Moran's I and Geary's C analysis by event type for the England dataset.
- **[Ramapuram Sreenivasa Raju] (IMT2023002):** Spatial heterogeneity analysis including coefficient of variation computation, Levene's test across pitch halves and thirds, and one-way ANOVA across defensive, midfield, and attacking thirds for all leagues. Spatial regression modelling OLS baseline, Spatial Lag (GM_Lag), and Spatial Error (GM_Error) estimation, residual Moran's I diagnostics, and MAE computation. GWR case study on Spain including AICc bandwidth selection and local coefficient mapping.

Acknowledgments

We would like to express our heartfelt gratitude to Professor Jaya Sreevalsan Nair for their exceptional guidance and teach-

ings in the SpatioTemporal Data Analytics course at IIITB. We are deeply grateful for the opportunity to undertake this assignment under their mentorship. We also acknowledge the use of ChatGPT and Gemini AI for obtaining information and generating segments of text and code; however, all analysis, interpretations, and intellectual contributions were solely our own, driven by the knowledge and skills acquired during the course.

REFERENCES

[1] A. Espinosa, "Soccer Match Event Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/aleespinosa/soccer-match-event-dataset