Spatial and Positional Effects on Defensive Events in the NBA

Noel Mamo, Savannah Gibbs, Pratham Lakhani, and Ryan Mulcahy

Introduction

Play styles in the NBA have evolved, and thus defensive scheming has been forced to evolve with it. The 5 best team offensive rating seasons off all time occured in the last two season, the team 3PA record was broken this year, and generally offenses are getting more and more efficient, farther and farther from the basket. The traditional consensus says that interior defense and rim protection are far and away the most important and effective defensive roles, but as the game adapts, we are wondering if that has changed. Where on the court are most defensive events happening. Who is accumulatingPut simply, we want to ask: **How do court location and player role impact defensive contribution in the NBA?** Specifically, we are looking for relationships between position groups, court location, and defensive events as we have defined and shaped around our data.

Data

Description

The dataset we used comes from the play by play data in the hoopR package. The dataset is made up of logged events with features including a type label, a text description, involved player ID's, and x-y coordinates for event location. We used specifically data from the 2023-24 NBA season as this seemed most relevant for assessing current trends. Before filtering and pre-processing, there were 614,447 observations with a total of 63 features.

Pre-processing

Identifying defensive events proved to be a challenge given the available data. There is a feature for a secondary and tertiary involved player for each event, but defensive actors are only included when a defensive box score is logged (steal, block, etc). To compensate for this we made the strong assumption that the offensive player logged for an event in the data was being guarded by a defensive player of the same position group. Thus, anything we classified as a defensive event (turnover, offensive foul, etc) we are interpreting as a defensive contribution for a player of the position of group of the offensive player listed in the data.

We defined 3 defensive events: **Possession Changes**, **Missed Shots**, and **Defensive Fouls**. Within our available data, these seemed like the most relevant way to categorize the existing events. We believe these are effective because they encompass most of the available event data, and define both positive and defensive events, which could be useful for inference later.

Because of the strong assumption regarding defensive player position mention earlier, we broadened positional categories from the 5 traditional positions (PG, SG, SF, PF, C) to the 3 general position groups (G, F, C). The positional data was joined from the box score data in the hoopR package using matching player ID's. After all filtering and preprocessing, we were left with 297,303 variables.

Exploratory Data Analysis

To visualize our data, and understand general trends in defensive events, we made the scatterplots below. In Figure 1, we made a spatial scatter plot of possession changing events. We discovered that defensive rebounds are the a frequent subtype, with most happening in a broad area around the basket, and right behind the three point line. Turnovers were the also very common with no discernible spacial pattern. Offensive fouls were the least common again with no discernible spacial pattern.

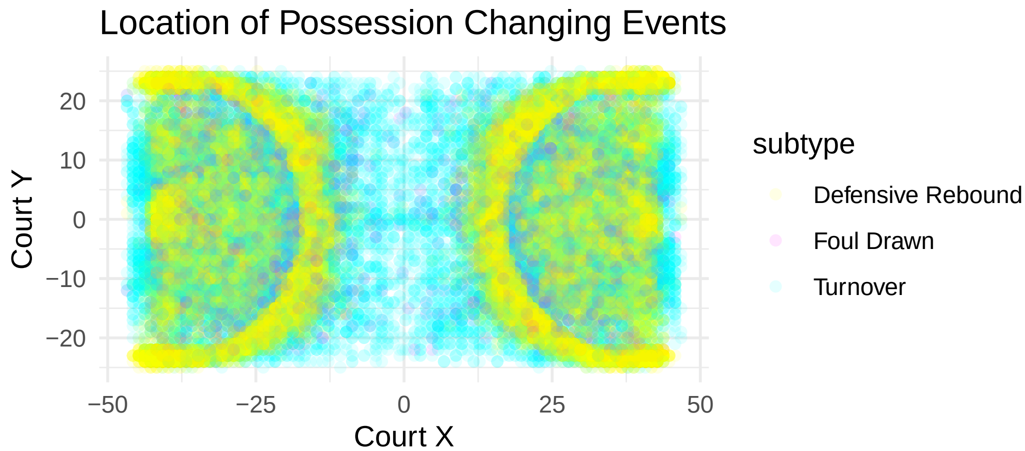

Figure 1: Spatial Plot of Possession Changing Events

In Figure 2, we made the same plot but for Missed Shot Events. The coloring display shot type, which is expectedly (and determinate in the case of the three point shots) strongly correlated with court location. This effectively just gives a reduced shot chart, and tells us more about where players shoot in general than where they specifically miss shots.

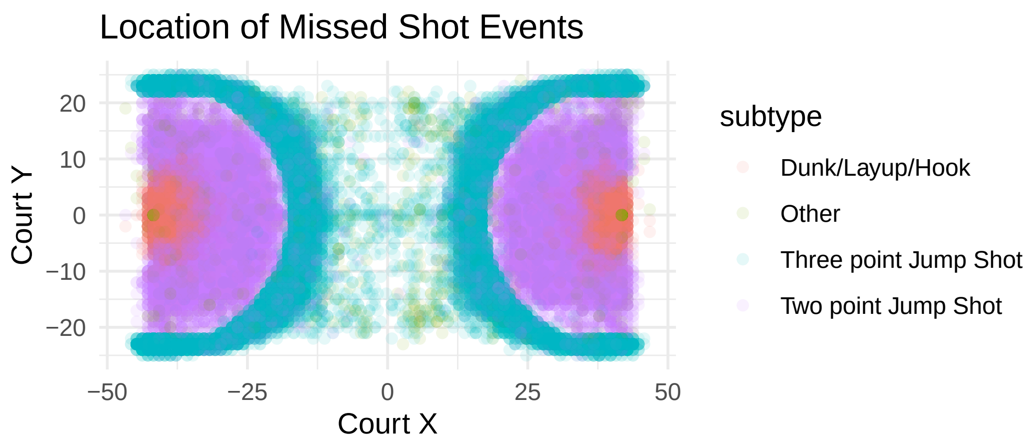

Figure 2: Spatial Plot of Missed Shot Events

Lastly in Figure 3, we make another spatial plot for Defensive Fouls. Here we are looking to see where defensive failures occur, and whether or not they lead directly to point opportunities. The free throw labeling is mainly here for EDA, but the feature could be used in the future for more detailed assessments about how these events lead to points.

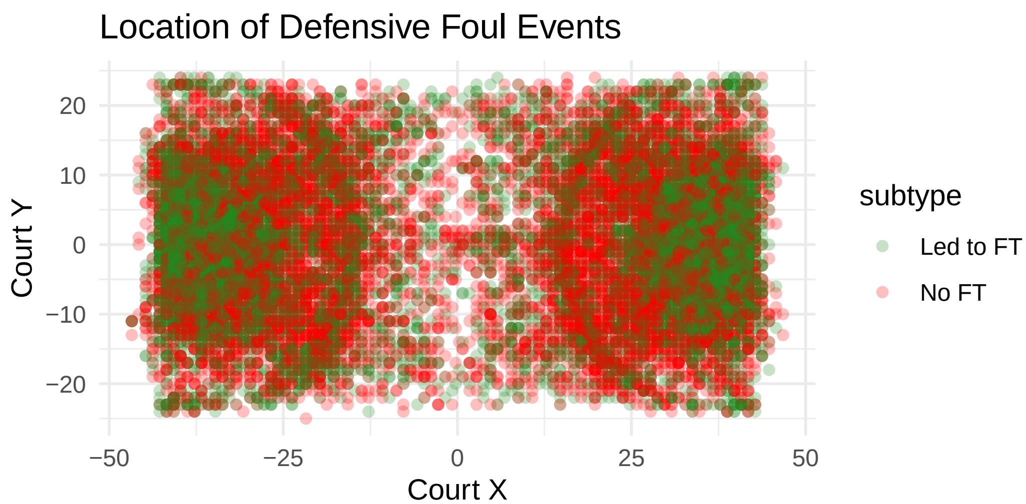

Figure 3: Spatial Plot of Defensive Foul Events

Methods

The goal of this project is to understand how player position group and spatial location on the court influence the type of defensive event that occurs in the NBA. To model this relationship, we used a multinomial logistic regression. The outcome variable is the type of defensive event, classified as one of three categories: foul, missed shot, or possession change. The explanatory variables are player position group (guard, forward, center) and spatial features derived from play-by-play event coordinates (euclidean distance and angle to the hoop).

Multinomial logistic regression is appropriate for this setting because the response is a categorical variable with three unordered outcomes. To satisfy the model's assumptions, including the requirement for monotonicity in the relationship between predictors and outcome log-odds, we transformed the original court coordinates into distance and angular features. Euclidean distance measures how far an event occurred from the basket, while angle to hoop captures directional differences across the court, allowing the model to differentiate, for example, plays occurring along the sidelines from those near the top of the key.

Model evaluation is based on comparisons using Akaike Information Criterion (AIC), which balances model fit and complexity. Uncertainty in the model coefficients is quantified using standard errors calculated through maximum likelihood estimation. This approach is appropriate given the large sample size and the asymptotic properties of the multinomial logistic model. We lastly evaluated our model using a confusion matrix, checking our model's predictive capability on unseen observations.

Results

Comparison

The models tested have type (type of defensive event) as their response variable and different permutations of euclidean distance, angle_to_hoop, and position group (with and without interactions) to find the model with the lowest AIC score. The model 'multinom_model' has the lowest score as seen in the table below, and thus is the model we moved forward with.

- multinom_model: type ~ position_group * euclid_dist + angle_to_hoop
- multinom_model_2: type ~ position_group + euclid_dist*angle_to_hoop
- multinom_model_3: type ~ position_group + euclid_dist + angle_to_hoop
- multinom_model_4: type ~ position_group + euclid_dist
- multinom_model_5: type ~ position_group * euclid_dist

AIC Comparison of Multinomial Models

Model	df	AIC
multinom_model	14	538,959.9
multinom_model_2	12	540,289.8
multinom_model_3	10	540,324.8
multinom_model_4	8	540,359.9
multinom_model_5	12	538,992.2

Model Results

The graphs below visualize the probabilities of defensive events occurring across a court. To maintain the monotonic assumption of the model, we had to convert coordinate values to euclidian distance from the basket and angle to the basket since having two opposite reference points was incompatible with our model. Taking this into account, we limit our interpretations of spatial results to distance and angle from the basket.

The heatmap in Figure 4 shows that turnovers and defensive stops are most likely to occur closer to the sidelines and farther away from the basket. Probability is lower near the center of the court and increases toward the baselines and sidelines, reflecting a higher risk of losing possession in pressured or transition heavy areas. This spatial pattern aligns can be interpreted with many common inferences from the game such as there being more possession changes in areas where you are less likely to shoot. Another potential explanation could be that passes travel farther in more open areas of the court (away from the baskets) and thus give longer opportunities for defensive disruptions.

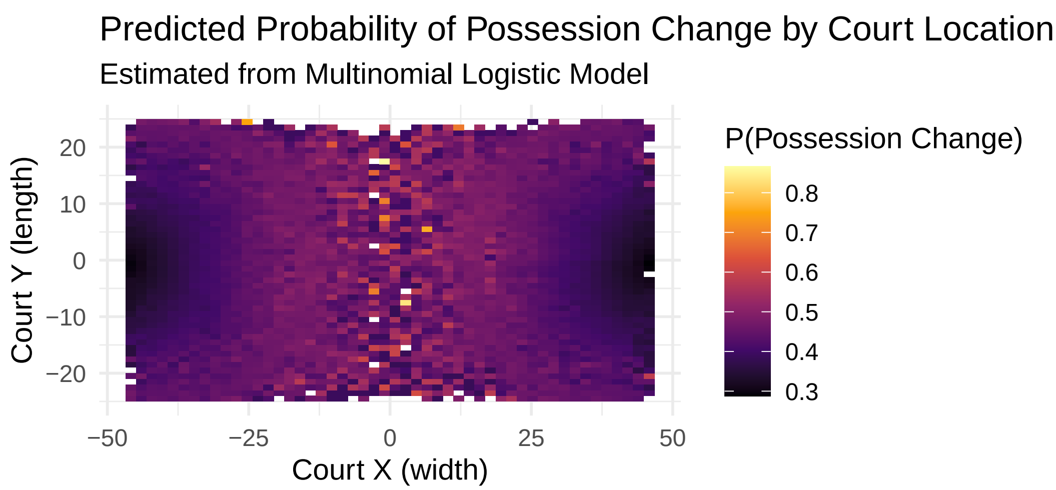

Figure 4: Predicted Probability of Possession Change by Court Location

In Figure 5, we can see that court location has a much weaker correlation to the Missed Shot event. There is a relatively uniform distribution across the court, with slightly lower probabilities near the baselines particularly right under the baskets. Missed shots appear somewhat more likely to occur in central areas closer to the basket and the top of the key, in line with the negative perception of contested midrange and interior attempts. This pattern suggests that while missed shots can happen throughout the court, there is very slight spatial concentration near particular scoring zones.

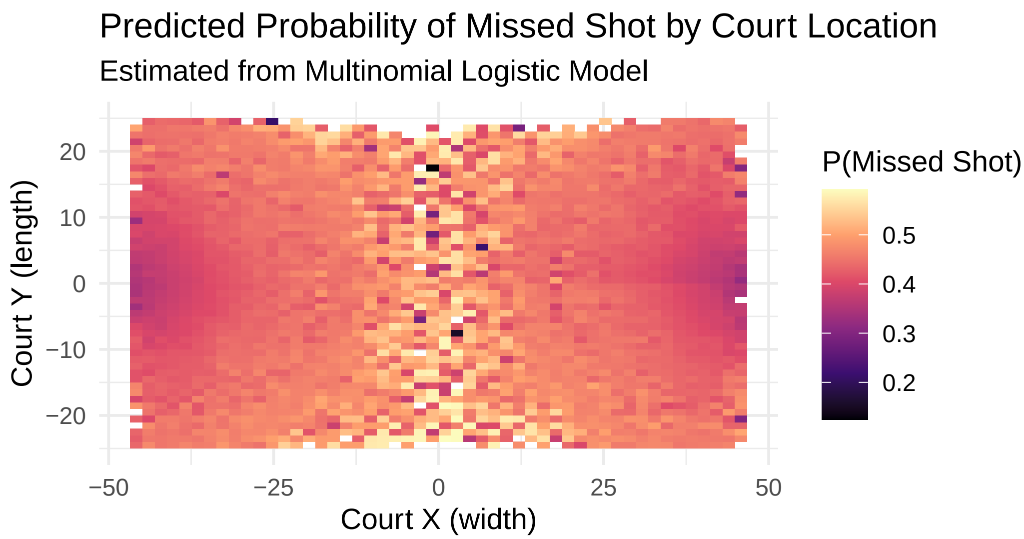

Figure 5: Predicted Probability of Missed Shot by Court Location

The heatmap of predicted foul probability in Figure 6 shows that fouls are most likely to occur near the baskets, with probabilities peaking close to each baseline and decreasing steadily toward center court. The smoothness of the plot could be due to the smaller sample size of this particular event. This pattern could reflect greater physical contact and defensive pressure in congested areas near the rim. In contrast, foul probability is lowest near midcourt, consistent with less contested action occurring farther from the paint.

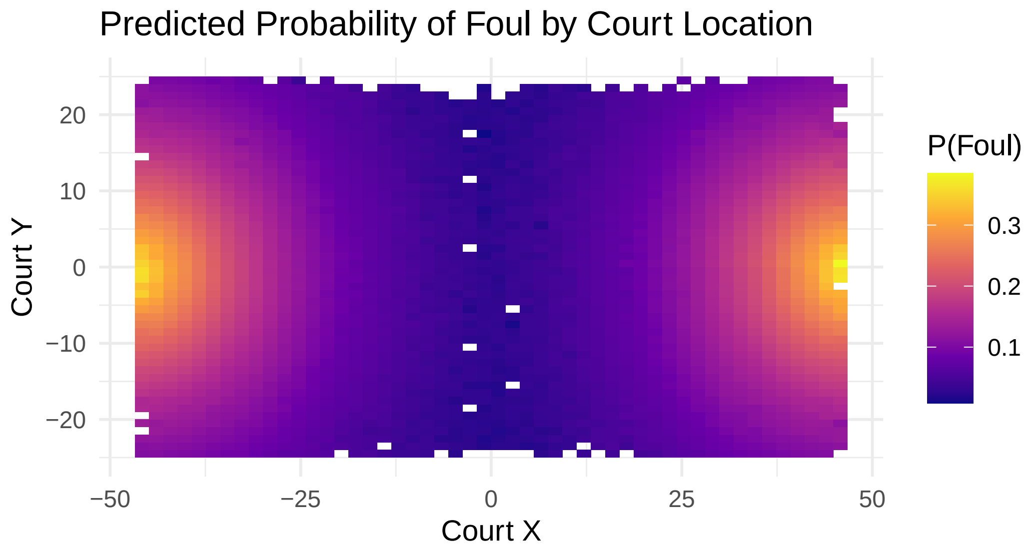

Figure 6: Predicted Probability of Foul by Court Location

Our fitted multinomial logistic regression model shows how player position group, Euclidean distance to the nearest basket, and angle relative to the hoop influence the likelihood of each defensive event type. In general, increased distance from the basket is associated with a higher probability of both missed shots and possession changes, particularly for players classified as forwards or guards. The interaction terms suggest that as distance increases, the likelihood of possession changes decreases slightly more for perimeter players compared to centers. The angle to the hoop has a small negative effect on the probability of a missed shot and a near zero effect on possession changes, indicating that directional differences have very limited influence. Overall, the model captures meaningful spatial and position based structure for defensive events.

Multinomial Model Coefficients

Event Type	Predictor	Coefficient	Standard Error
Missed Shot	(Intercept)	-0.232	0.030
Missed Shot	Position: Forward	0.099	0.035
Missed Shot	Position: Guard	0.264	0.035
Missed Shot	Euclidean Distance	0.059	0.002
Missed Shot	Angle to Hoop	-0.011	0.003
Possession Change	(Intercept)	-0.391	0.029

Possession Change	Position: Forward	0.097	0.034
Possession Change	Position: Guard	0.222	0.034
Possession Change	Euclidean Distance	0.101	0.002
Possession Change	Angle to Hoop	0.002	0.003

Evaluation

To evaluate model performance, we used a confusion matrix comparing predicted event types to actual observed outcomes. The model achieved an overall classification accuracy of 47.3%, which exceeds the baseline no information rate of 42.8% (p value near 0). The model was most accurate in predicting missed shots, with a sensitivity of 67% and a balanced accuracy of 54%. It performed moderately well in predicting possession changes, but showed limited ability to detect fouls, likely due to the relatively low number of samples in the dataset. These results suggest that the model captures some useful spatial and positional structure, particularly for the higher frequency defensive events.

Confusion Matrix: Predicted vs Actual Event Type

Prediction	Foul	Missed Shot	Possession Change
Foul	6	5	25
Missed Shot	31,063	79,192	63,267
Possession Change	11,873	38,354	50,749

Discussion

This project set out to understand how player position group and court location influence defensive events occuring in the NBA. Using a multinomial logistic regression model incorporating position group, Euclidean distance to the basket, and angle to the hoop, we found that spatial features meaningfully contribute to predicting whether a defensive event will result in a foul, missed shot, or possession change. In particular, the model showed that possession changes were most likely along the sidelines, fouls were concentrated near the basket, and missed shots were more evenly distributed but slightly more common near key scoring zones. The model achieved an overall classification accuracy of 47.3%, outperforming a baseline no-information rate, suggesting that spatial and positional features provide useful, though imperfect, predictive signal.

There are several limitations to the modeling approach used. First, multinomial logistic regression assumes a linear relationship between the predictors and the log-odds of each outcome, which may not fully capture the complexity of basketball spatial dynamics. Although the transformations we made to the coordinate features helped address the basic monotonic requirements, they may not represent all nuanced aspects of play, such as defender positioning, preventative actions that won't show up as an event, or the flow of possessions. Additionally, the model struggled to predict less frequent events like fouls accurately, likely due to the imbalance of sample size in the response classes. More sophisticated methods, such as Bayesian hierarchical models or models incorporating real time tracking data could potentially address these shortcomings.

Future Work: Bayesian Multilevel Modeling

Reapproaching this project with a Bayesian multilevel model could offer several advantages, particularly for handling hierarchical structure, partial pooling, and improved understanding of spatial uncertainty. We could more rigorously account for position level differences, and better quantify uncertainty with regards to our spacial effects. This approach could allow us to create a more interpretable, more precise, and more flexible model and improve the impact of this project in the future.