Trabajo Fin de Máster

Estudio de patrones de lanzamiento en béisbol mediante técnicas no paramétricas

Natanael Ventura Jiménez

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo dos padróns de lanzamento no béisbol mediante técnicas non paramétricas
Título en español: Estudio de patrones de lanzamiento en béisbol mediante técnicas no paramétricas
English title: Study of pitching patterns in baseball using non-parametric techniques
Modalidad: Modalidad A
Autor/a: Natanael Ventura Jiménez, Universidad de Vigo
Director/a: María Alonso Pena, Universidad de Santiago de Compostela; Jose Ameijeiras Alonso, Universidad de Santiago de Compostela
Breve resumen del trabajo: El estudio analiza los patrones de lanzamiento en béisbol usando técnicas no paramétricas. Se utilizaron datos de la temporada regular de Grandes Ligas 2024. La metodología aplicada incluye modelos de regresión spline aditiva para identificar cómo la ubicación y el desplazamiento de los lanzamientos influyen en la velocidad y el ángulo vertical de salida de la pelota al ser bateada. En general, los resultados muestran que los lanzamientos localizados en las esquinas inferiores de la zona de bateo tienden a generar contactos más débiles, y que el movimiento vertical de los lanzamientos es crucial para inducir batazos rodados. Estos hallazgos proporcionan información valiosa para desarrollar estrategias de lanzamientos más efectivas.

Doña María Alonso Pena, Profesora ayudante de la Universidad de Santiago de Compostela, don Jose Ameijeiras Alonso, Profesor permanente de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Estudio de patrones de lanzamiento en béisbol mediante técnicas no paramétricas

fue realizado bajo su dirección por don Natanael Ventura Jiménez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Vigo, a 12 de enero de 2025.

La directora:
Doña María Alonso Pena

Firmado por
AMEIJEIRAS
ALONSO JOSE -

El director:
Don Jose Ameijeiras Alonso

Firmado por
VENTURA JIMENEZ
NATANAEL -

El autor:
Don Natanael Ventura Jiménez

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen

Resumen en español

El estudio analiza los patrones de lanzamiento en béisbol usando técnicas no paramétricas. Se utilizaron datos de la temporada regular de Grandes Ligas 2024. La metodología aplicada incluye modelos de regresión spline aditiva para identificar cómo la ubicación y el desplazamiento de los lanzamientos influyen en la velocidad y el ángulo vertical de salida de la pelota al ser bateada. En general, los resultados muestran que los lanzamientos localizados en las esquinas inferiores de la zona de bateo tienden a generar contactos más débiles, y que el movimiento vertical de los lanzamientos es crucial para inducir batazos rodados. Estos hallazgos proporcionan información valiosa para desarrollar estrategias de lanzamientos más efectivas.

English abstract

The study analyzes pitching patterns in baseball using non-parametric techniques. Data from the 2024 MLB regular season were used. The methodology includes additive spline regression models to identify how pitch location and movement influence the ball's exit velocity and vertical launch angle when hit. Overall, the results show that pitches located in the lower corners of the plate tend to generate weaker contact, and vertical pitch movement is crucial for inducing ground balls. These findings provide valuable insights for developing more effective pitching strategies.

Capítulo 1

Introducción

El béisbol es un deporte de equipo con una arraigada tradición en América y una creciente popularidad en Asia. En Estados Unidos, es considerado el pasatiempo nacional y su influencia se extiende a numerosos países latinoamericanos como República Dominicana, Panamá, Cuba, Curazao, Aruba, Nicaragua, Puerto Rico, México y Venezuela, donde es uno de los deportes más practicados. A su vez, en Asia, el béisbol goza de gran popularidad en Taiwán, Corea del Sur y Japón. Las Grandes Ligas de Béisbol (MLB) es la principal organización de béisbol profesional del mundo, reuniendo a 30 equipos de Norteamérica y a jugadores de diversas nacionalidades, especialmente de Latinoamérica y Asia.

En un partido de béisbol, dos equipos de nueve integrantes se alternan a la ofensiva y defensiva durante nueve encuentros o entradas. El objetivo es anotar más carreras que el oponente. Los bateadores buscan conectar la pelota y avanzar por las bases, mientras que los lanzadores y el resto de los jugadores defensivos intentan eliminarlos. Al batear la pelota, los jugadores pueden lograr diferentes resultados dependiendo del número de bases que logren alcanzar (sencillo, doble, triple o cuadrangular). La habilidad de los bateadores y la estrategia de los lanzadores inciden en el resultado de cada jugada. Los lanzadores manipulan las características (velocidad, localización y desplazamiento) y la secuenciación de sus lanzamientos con el objetivo de dificultar que sean conectados con éxito por los bateadores rivales.

El béisbol es un deporte altamente cuantificable, con una gran cantidad de estadísticas que permiten analizar el desempeño de cada jugador y equipo con precisión. A finales de la década de 1970, surgió el término sabermetría para referirse a la disciplina que busca entender el juego a través del análisis empírico de los datos. Desde métricas tradicionales como el promedio de bateo y las carreras impulsadas, hasta estadísticas avanzadas como el WAR (Wins Above Replacement), los equipos han utilizado una vasta cantidad de datos para tomar decisiones estratégicas. En la MLB, esta tendencia se ha acelerado en años recientes, impulsada por los avances tecnológicos. Hoy en día, cada lanzamiento, cada bateo y cada movimiento en el campo puede ser medido, analizado y desglosado en múltiples estadísticas.

1.1. Problema de investigación

Dado que el béisbol gira en torno a la interacción entre lanzadores y bateadores, resulta esencial investigar los factores que influyen en los resultados de cada enfrentamiento y extraer los patrones subyacentes. Algunas cuestiones de interés podrían ser: ¿Qué características de un lanzamiento hacen que sea más difícil de conectar? o ¿Qué características comparten los lanzamientos que generan ciertos tipos de contactos por parte del bateador? La respuesta a estas preguntas permitiría a los jugadores y equipos desarrollar estrategias más efectivas. El presente trabajo se centra en estudiar la relación

entre las características de los lanzamientos y los resultados de los contactos. En general, el objetivo del estudio es caracterizar la relación de dependencia entre las características de los lanzamientos y los resultados de los contactos de bateo. De forma concreta, los objetivos del estudio son:

1. Identificar las características distintivas de los lanzamientos que producen contactos de bateo con velocidades de salida extremas (débil/fuerte).

2. Determinar las características distintivas de los lanzamientos asociados a contactos con distintos ángulos verticales de salida (rodados, líneas o elevados).

En lugar de caracterizar las observaciones individuales X_i asociadas a valores específicos de Y_i, lo cual se conseguiría mediante un análisis descriptivo, sería más útil analizar los valores de X_i para los cuales la esperanza condicional de Y alcanza determinados niveles.De esta manera, se puede aislar mejor el efecto de las características de los lanzamientos en el resultado del bateo.

Para superar las limitaciones del análisis descriptivo y modelar la compleja relación entre las variables de lanzamiento y los resultados de bateo, se propone un enfoque de regresión no paramétrica. Este método permite flexibilizar la forma funcional de la relación, adaptándose a patrones no lineales que podrían no ser capturados por modelos paramétricos tradicionales. En este modelo, la velocidad y el ángulo de salida del bateo se expresarán como función de un conjunto de covariables que incluyen las características del lanzamiento (velocidad, ubicación y movimiento) y variables de control relevantes [1]. El análisis de los conjuntos de nivel de la función de regresión estimada permite la extracción de patrones en los lanzamientos que dan lugar a tipos de contactos de bateo específicos.

1.2. Antecedentes

Hasta hace poco, la escasez de datos desagregados limitaba los análisis de enfrentamientos en béisbol, los cuales se basaban en estadísticas agregadas como proporciones y promedios. Modelos como los logísticos se empleaban comúnmente para predecir resultados como hits, rodados o embasamientos, utilizando tasas de ponches tanto individuales como de la liga (Healey, 2017, Healey, 2015, Doo and Kim, 2018). Sin embargo, este enfoque, aunque útil, ofrece una visión limitada de los factores específicos que influyen en el desempeño de un bateador en cada turno. Otros enfoques, si bien también se sustentan en datos agregados, profundizan en variables como las características de los lanzamientos, el contexto del partido y la incertidumbre, proporcionando una comprensión más rica de los factores que determinan el éxito o fracaso de un bateador.

Respecto a las características de los lanzamientos, múltiples estudios han demostrado su influencia en el éxito de los bateadores. Por ejemplo, Nakahara et al. (2023), mediante un análisis estratificado basado en puntaje de propensión, encontró que la estrategia de lanzar fuera es más efectiva que la de lanzar dentro, incluso al controlar por las habilidades del bateador y las preferencias de localización del lanzador. Asimismo, Yee and Deshpande (2024), utilizando árboles de regresión aditivos bayesianos, determinó que la ubicación del lanzamiento es el factor más importante para predecir si el bateador hará contacto con la pelota. A su vez, Healey (2019), utilizando regresión no paramétrica, predice el desempeño esperado de cada lanzamiento en términos del promedio de embasarse ponderado (wOBA). Utilizando como variables explicativas, entre otros factores, la velocidad, localización y desplazamiento de cada lanzamiento, determina que la capacidad predictiva del modelo supera otras métricas tradicionales.

Diversos estudios han evidenciado la influencia del contexto del partido en el desempeño de los bateadores y lanzadores. Por ejemplo, Healey (2019) y Yee and Deshpande (2024) han demostrado que

[1] En el estudio se controla por factores externos a cada lanzamiento como: conteo de bolas-strikes, lanzamientos previos, entrada y corredores en base.

factores como la entrada, los corredores en base y el conteo de bolas y strikes influyen significativamente en el resultado de cada turno al bate. De hecho, Gray (2002), en un experimento controlado, encontró que los bateadores tienen una ventaja clara cuando están delante en el conteo, lo que sugiere que adaptan su estrategia de bateo según el contexto. Este comportamiento adaptativo no se limita a los bateadores, ya que Cox et al. (2017) demostró que los lanzadores también ajustan su repertorio de lanzamientos a lo largo del partido, según lo revela un análisis de los patrones de selección de lanzamientos de cinco abridores de la MLB en 2014.

La incertidumbre que enfrenta el bateador respecto a las características de los lanzamientos también influye significativamente en los resultados. Además de la variabilidad en velocidad, localización y movimiento, la secuenciación de los lanzamientos introduce un elemento adicional de incertidumbre. Martin (2019), utilizando árboles de decisión, demostró que variaciones en el desplazamiento vertical de los lanzamientos están fuertemente asociadas con tasas de ponche más altas. Asimismo, mediante un análisis de regresión lineal, Healey and Zhao (2017) encontró que la correlación entre las características de lanzamientos consecutivos impacta negativamente en la tasa de ponche de los lanzadores. Por su parte, Kim and Jung (2018) utilizó la información mutua normalizada entre tipo de lanzamiento y localización para cuantificar la incertidumbre en el repertorio de lanzadores de MLB. El estudio no es concluyente sobre la relación entre el grado de incertidumbre y el desempeño general de los lanzadores, no obstante, indica que los lanzadores de élite suelen destacar en al menos una de estas dimensiones: incertidumbre lanzamiento-localización o velocidad.

1.3. Estructura del documento

A continuación, se detalla la estructura del trabajo. El segundo capítulo se describe la base de datos utilizada, se definen las variables relevantes y se realiza un análisis exploratorio de los datos, poniendo especial énfasis en la relación entre los lanzamientos y los batazos, así como en el efecto del tipo de enfrentamiento. En el tercer capítulo se realiza una revisión bibliográfica de diferentes técnicas de regresión no paramétrica y otras metodologías que se utilizarán para la detección de patrones en los lanzamientos. Los resultados obtenidos a partir de los análisis realizados se presentan en el cuarto capítulo. Finalmente, en el quinto capítulo se presentan las conclusiones del estudio, resumiendo los principales hallazgos y discutiendo sus implicaciones.

Para garantizar la claridad y precisión en la comunicación, se ha incorporado al final del documento un glosario que define los términos técnicos, en su mayoría anglicismos, utilizados en el análisis. Esta herramienta facilita la comprensión del trabajo, especialmente aquellos términos que no cuentan con una traducción directa al castellano o que presentan múltiples significados en el contexto del béisbol.

Capítulo 2

Descripción y exploración de los datos

En este capítulo se presenta un análisis descriptivo de los datos de Statcast con el objetivo de explorar la relación entre las características de los lanzamientos y los resultados de los batazos, así como el impacto del tipo de enfrentamiento lanzador-bateador. A través de un análisis detallado de las variables relacionadas con los lanzamientos y los contactos de bateo, se ha identificado una relación estadística significativa entre ambas. Los resultados obtenidos sugieren que existen diferencias significativas en las características de los lanzamientos según el tipo de lanzamiento, el tipo de contacto y el tipo de enfrentamiento, lo que indica que estas variables están relacionadas de manera compleja.

2.1. Sistema statcast

Statcast es un sistema de seguimiento de última generación implementado por la MLB para cuantificar de manera precisa una amplia gama de acciones durante los partidos. Sus orígenes se remontan a 2008, con la instalación inicial de dispositivos de seguimiento de lanzamientos. Sin embargo, fue a partir de 2015 que el sistema alcanzó madurez, con la implementación de tecnología más avanzada en todos los estadios. Este desarrollo permitio un seguimiento detallado de múltiples aspectos del juego, antes inaccesibles a la medición objetiva.

Entre 2015 y 2019, Statcast empleaba una combinación de sistemas ópticos estereoscópicos y radar para el seguimiento de la pelota y los jugadores (Mizels et al., 2022). Sin embargo, a partir de 2020, la incorporación del sistema Hawk-Eye mejoró considerablemente la precisión y el alcance de las métricas obtenidas. Este nuevo sistema, equipado con doce cámaras de alta velocidad, permite un seguimiento más detallado y preciso de cada acción en el campo de juego. Cinco de las cámaras del sistema se encargan exclusivamente del seguimiento de los lanzamientos y el bate, mientras que el resto de las cámaras realiza el seguimiento de las personas en el terreno (jugadores y árbitros) y las pelotas bateadas (McElroy) . Gracias a Hawk-Eye, Statcast ofrece ahora capacidades avanzadas como el análisis biomecánico de los jugadores, la cuantificación de la velocidad y trayectoria del swing, y una medición precisa de las características de cada lanzamiento. Estas mejoras han revolucionado la comprensión del béisbol, permitiendo un análisis más profundo del rendimiento de los jugadores y el desarrollo de nuevas estrategias.

2.2. Definición de variables

El sistema Statcast clasifica los lanzamientos en tres categorías principales: Fastball, Offspeed y Breaking (ver Anexo 1). Los lanzamientos del primer grupo, recta (fastball), se caracterizan por su alta velocidad y mínimo desplazamiento. En el grupo Offspeed (lanzamientos de velocidad reducida) se encuentran aquellos de baja velocidad con un desplazamiento vertical abrupto. Por otro lado, el grupo rompientes (breaking) incluye lanzamientos con un mayor grado de desplazamiento vertical y horizontal, como curvas y sliders. Aunque esta clasificación inicial es útil, no captura completamente la diversidad de movimientos y velocidades de cada lanzamiento. Por ello, para caracterizar detalladamente cada lanzamiento, se recurre a una cuantificación directa de variables como velocidad, localización y desplazamiento (ver Cuadro 2.1).

El resultado obtenido por un bateador al hacer contacto con la pelota puede caracterizarse de múltiples formas, siendo la velocidad de salida una de las métricas más importantes para evaluar la calidad del contacto. Una mayor velocidad de salida brinda menos tiempo de reacción a la defensa, incrementando las posibilidades de éxito. Asimismo, el ángulo vertical de salida proporciona información complementaria sobre la calidad del contacto (ver Cuadro 2.1).

Basados en el ángulo de salida de la pelota, los contactos se clasifican en: rodados ($< 10°$), líneas (10°-25°), elevados (25°-50°) y popup ($> 50°$). Los contactos clasificados como "líneas" y "elevados" tienen una mayor probabilidad de convertirse en *hits* debido a su trayectoria. Por lo tanto, el rango de ángulos entre 8° y 32° es indicador de un contacto de alta calidad y se le denomina "sweet spot", lo que sugiere un impacto en la parte óptima del bate.

La categoría más exclusiva es "barrel". Un "barrel" se define como un contacto con una velocidad de salida superior a 97 mph y un ángulo de salida entre 26° y 30°, aunque estos límites se ajustan ligeramente con cada incremento en la velocidad llegando a un máximo de 8° a 50° para velocidades de 116 mph o más. Los "barrels" representan los contactos de mayor calidad y tienen una alta probabilidad de generar extrabases.

2.3. Análisis descriptivo

Para llevar a cabo el análisis cuantitativo de la interacción entre los lanzamientos y los contactos de bateo, se emplea una base de datos construida a partir de los datos de Statcast correspondientes a la temporada regular de Grandes Ligas 2024 (20 de marzo - 30 de septiembre). Esta base se restringe a los lanzamientos conectados en zona de juego (fair ball), excluyendo los toques de sacrificio, enfrentamientos con lanzadores con menos de 50 lanzamientos en la temporada, así como los lanzamientos no clasificados o de tipos poco frecuentes (menos del 5 % del total). Asimismo, se eliminarán los datos correspondientes a enfrentamientos en los que el lanzador agotó su turno al bate.

Del total de 110.341 lanzamientos analizados, la recta de 4 costuras (31,2 %) es el lanzamiento más frecuente, seguida del sinker (20,3 %) y el slider (15,8 %). Cada tipo de lanzamiento exhibe patrones distintivos en términos de velocidad, ubicación y movimiento (Cuadro 2.3). La recta de cuatro costuras, con una velocidad promedio de 94,6 mph, es el lanzamiento más rápido. En contraste, la curva presenta la velocidad promedio más baja, de 79,6 mph. Respecto a la ubicación, si bien existe una tendencia general a ubicar los lanzamientos en la parte interna del plato, la recta de cuatro costuras y el sinker se desvían de este patrón, presentando un localización horizontal promedio ligeramente hacia la zona externa.

En cuanto al movimiento, tanto el sinker como el cambio y el sweeper exhiben el mayor desplazamiento horizontal, lo que indica una mayor tendencia a moverse hacia la zona externa del plato. Por

Cuadro 2.1: Descripción de variables seleccionadas

Variables	**Descripción**	**Unidad**
Velocidad	Velocidad efectiva del lanzamiento	Millas por hora (mph)
Localización horizontal (Lh)	Posición horizontal de la pelota cuando cruza en plato de bateo desde la perspectiva del receptor. Se mide con respecto a una línea vertical imaginaria justo en el centro del plato. En Statcast se asigna valores positivos o negativos según la pelota se encuentre a la derecha o izquierda de dicha línea. En el presente estudio se modifica la métrica de tal forma que tome valores positivos solo cuando la pelota y el bateador se encuentran del mismo lado de la línea vertical central.	pies
Localización vertical (Lv)	Posición vertical de la pelota cuando cruza en plato de bateo, se mide con respecto al suelo desde la perspectiva del receptor.	pies
Desplazamiento vertical (Dv)	Desplazamiento vertical de la pelota visto desde la perspectiva del receptor. Se calcula como la diferencia entre las coordenadas verticales de la pelota en el punto de liberación por el lanzador y el punto donde atraviesa el plato de bateo.	pies
Desplazamiento horizontal (Dh)	Desplazamiento horizontal de la pelota visto desde la perspectiva del receptor. Se calcula como la diferencia entre las coordenadas horizontales de la pelota en el punto de liberación por el lanzador y el punto donde atraviesa el plato de bateo. La métrica es positiva si el desplazamiento se da en dirección al bateador y negativa en caso contrario.	pies
Ángulo de salida (As)	Ángulo vertical (respecto al suelo) en el que la pelota sale del bate inmediatamente después del contacto.	Grados (°)
Velocidad de salida (Vs)	Velocidad a la que la pelota sale del bate inmediatamente después del contacto.	Millas por hora (mph)

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*

otro lado, la curva y el cambio presentan el mayor desplazamiento vertical, sugiriendo un movimiento pronunciado hacia abajo. Es importante destacar que, aunque la recta de cuatro costuras y el sinker tienden a moverse hacia la zona externa, su desplazamiento horizontal es menor en comparación con otros tipos de lanzamientos como el sinker, el cambio y el sweeper.

Como se observa en el Cuadro 2.4, en el 72,6 % de los enfrentamientos considerados el lanzador involucrado es distro, siendo los enfrentamientos contra bateadores diestros los más comunes (33,3 %). A su vez, los enfrentamientos entre lanzadores y bateadores zurdos son los menos comunes (7,9 %).

Cuadro 2.3: Resumen estadístico de lanzamientos; frecuencia, velocidad, localización y desplazamiento

Lanzamiento				Localización		Desplazamiento	
Grupo	**Nombre**	**Frecuencia Absoluta**	**Velocidad**	x	z	x	z
Recta	Recta de 4 costuras	34.420	94,557	0,034	2,683	1,822	3,151
Recta	Sinker	22.443	93,505	0,116	2,393	2,086	3,225
Rompiente	Slider	17.435	86,011	-0,170	2,176	1,840	3,596
Offspeed	Cambio	13.183	85,699	-0,358	2,079	1,908	3,648
Recta	Cutter	9.960	89,975	-0,007	2,449	1,727	3,414
Rompiente	Sweeper	6.552	81,888	-0,263	2,229	1,868	3,255
Rompiente	Curva	6.348	79,590	-0,164	2,082	1,730	3,855
Global		110.341	89,900	-0,061	2,390	1,880	3,370

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *El movimiento horizontal (z) se expresa en terminos absolutos. La velocidad se expresa en mph y las distancias en pies.*

Respecto a los contactos, el 39,6 % superó las 95 mph (contacto fuerte) y en el 35.3 % en ángulo de salida se situado entre 8 y 32 grados (sweet spot). No obstante, solo el 8,1 % de los contactos alcanzó la combinación óptima de velocidad y ángulo de salida (barrel). Además, los batazos rodados fueron los más comunes (43,58 %), seguidos por las líneas (23,5 %) y los elevados (23.0 %). En cuanto a los resultados de juego, el 67,7 % de los eventos resultó en out, mientras que el 20,9 % en sencillos, el 6,3 % en dobles y el 4,5 % en cuadrangulares.

Cuadro 2.4: Frecuencia absoluta y relativa de lanzamientos según tipo de enfrentamiento y características del contacto de bateo

Categorías	**Lanzamientos**	**%**
Tipo de enfrentamiento		
DD	43.205	39,30 %
DI	36.656	33,34 %
ID	21.409	19,47 %
II	8.674	7,89 %
Calidad del contacto		
Contacto fuerte	43.540	39,60 %
Sweet spot	38.830	35,32 %
barrel	8.898	8,09 %
Ángulo salida		
Rodado	47.915	43,58 %
Línea	25.813	23,48 %
Elevado	25.239	22,96 %
Popup	10.977	9.98 %
Resultado de contacto		
Out	74.386	67,66 %
Sencillo	23.044	20,96 %
Doble	6.915	6,29 %
Cuadrangular	4.979	4,53 %
Triple	620	0.56 %

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *La variable "tipo de enfrentamiento" contiene las categorias; DD, DI, II y ID, según la combinación de brazo de lanzar del lanzador y lado de bateo del bateador (D: diestro, I: zurdo). En la categorías Out se agragan los sacrificios, errores, outs forzados y jugadas de seleccion (fielder's choice).*

Antes de profundizar en el análisis de la relación entre las características de los lanzamientos y la calidad del contacto, es fundamental establecer la existencia de una relación estadística entre los conjuntos de variables que describen ambos eventos. Además, resulta esencial evaluar si esta relación

varía según el tipo de enfrentamiento entre lanzador y bateador.

2.3.1. Relación entre lanzamientos batazos

Una primera exploración de la relación entre lanzamientos y contactos se realizo a partir de tablas de contingencias. Al comparar la familia del tipo de lanzamiento (recta, offspeed o rompiente) con las indicadoras del tipo de contacto "barrel" o "contacto fuerte" ($>$ 95 mph) la prueba χ^2 cuadrada de Pearson arroja p-valores muy bajos (ver Anexo 2), indicando la existencia de una relación de dependencia. Un resultado similar se obtienes al reemplazar las indicadoras de tipo de contacto por la categorización del ángulo de salida (rodado, línea, elevado, popup). Estos resultados sugieren que el tipo de lanzamiento influye de manera determinante en el tipo de contacto logrado por el bateador. Asimismo, se empleó la prueba de rangos de Wilcoxon (ver Cuadro 2.5) para comparar las distribuciones de la velocidad, desplazamiento y localización (vertical y horizontal) de los lanzamientos en función del tipo de contacto del bateador. Los resultados muestran que las distribuciones de velocidad, desplazamiento horizontal y vertical de los lanzamientos difieren significativamente entre contactos clasificados como "barre" y el resto. En el caso de "contactos fuertes", aunque se observan diferencias significativas la mayoría de las variables, la hipótesis de igualdad en las distribuciones de localización horizontal y desplazamiento vertical no puede ser rechazada de manera contundente.

Cuadro 2.5: Media, diferencia en media contrafactual y p-valor correspondientes a la prueba de rangos Wilcoxon sobre igualdad en distribución de variables seleccionadas frente "barrel" y "contacto fuerte"

Variable	**barrel**		**contacto fuerte**	
	Media	p-valor	Media	p-valor
Velocidad	90,183 (0,300)	8,624e-06	90,581 (1,116)	2,20e-16
Localización horizontal	0,003 (0,070)	2,2e-16	-0,058 (0,007)	0,073
Localización vertical	2,477 (0,096)	2,2e-16	2,398 (0,016)	0,007
Desplazamiento horizontal	0,266(0,134)	8,16e-11	0,178 (0,065)	2,295e-07
Desplazamiento vertical	3.299 (-0,073)	2,20e-16	3,360 (-0,012)	0,063

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas *Diferencia en media contrafactual entre paretesis, Se refiere a la diferencia en media entre los batazos clasificados como "barrel" y "contacto fuerte" y los que no alcanza esa clasificación. .*

Adicionalmente, se estudió cómo varían la velocidad, localización y desplazamientos (vertical y horizontal) de los lanzamientos según el tipo de contacto definido por el ángulo de salida de la pelota (rodado, línea, elevado o popup). Los resultados de la prueba de Kruskal-Wallis (ver Anexo 3) mostraron heterogeneidad en las distribuciones de las variables de los lanzamientos entre las diferentes categorías de tipo contacto. Se realizaron comparaciones múltiples de Nemenyi para identificar las diferencias específicas entre grupos.

Los resultados de la prueba de Nemenyi (ver Cuadro 2.6) indican que la distribución de la velocidad de los lanzamientos es estadísticamente similar en los pares de rodados y líneas, líneas y pop-up, así como elevados y pop-up, pero difiere significativamente en el resto de las combinaciones posibles. Además, las distribuciones de la localización vertical y horizontal de los lanzamientos son estadísticamente distintas para cada uno de los resultados de bateo considerados.

En cuanto a la distribución del desplazamiento horizontal de los lanzamientos, no se observan diferencias significativas entre los grupos formados por líneas y pop-up, ni entre pop-up y rodados, mientras que para el resto de combinaciones de resultados, la distribución del desplazamiento horizontal es estadísticamente diferente. De manera similar, la distribución del desplazamiento vertical de los lanzamientos no presenta similitud entre ninguna de las posibles combinaciones de resultados de bateo.

En resumen, los resultados obtenidos indican que, en general, existen diferencias significativas en las características de los lanzamientos según el ángulo de vertival salida de la pelota.

Cuadro 2.6: Medias y códigos de significacción correspondientes a la prueba de de rangos múltiples de Nemenyi sobre la igualdad en distribución de las variables de lanzamientos entre los distintas clasificaciones del "ángulo vertical de salida".

Grupos	**Medias**	**Grupos**		
		Rodados	Lineas	Elevados
	Velocidad			
Rodados	90,058			
Líneas	89,928			
Elevados	89,640	***	***	
Popup	89,812	**		
	Localización horizontal			
Rodados	-0,112			
Líneas	-0,070	***		
Elevados	-0,012	***	***	
Popup	0,081	***	***	***
	Localización vertical			
Rodados	2,274			
Líneas	2,392	***		
Elevados	2,497	***	***	
Popup	2,630	***	***	***
	Movimiento horizontal			
Rodados	0,056			
Líneas	0,176	***		
Elevados	0,255	***	***	
Popup	0,142			***
	Movimiento vertical			
Rodados	3,461			
Líneas	3,372	***		
Elevados	3,277	***	***	
Popup	3,150	***	***	***

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *Código de significancia estadísticas:* $\alpha < 0,001 = ***$, $\alpha < 0,01 = **$, $\alpha < 0,05 = *$.

2.3.2. Efecto del tipo de enfrentamiento

Primero se evaluó si existen diferencias significativas en las características de los lanzamientos según el tipo de enfrentamiento lanzador-bateador. Para ello, se utilizó la prueba no paramétrica de Kruskal-Wallis (Anexo 4), la cual reveló que las distribuciones de las variables que caracterizan los lanzamientos (velocidad, localización y desplazamiento) difieren significativamente en al menos uno de los cuatro tipos de enfrentamientos posibles. A fin de identificar los tipos de enfrentamientos entre los que existen diferencias se aplicó la prueba de rangos múltiples de Nemenyi.

En el Cuadro 2.7 detalla los resultados de la prueba de Nemenyi, revelando patrones interesantes en las características de los lanzamientos según el tipo de enfrentamiento. Si bien la velocidad de los lanzamientos varía significativamente entre los diferentes escenarios, las demás variables muestran resultados más heterogéneos. La ubicación horizontal de los lanzamientos de los zurdos se diferencia estadísticamente del resto. En cuanto a la localización vertical, los lanzadores diestros muestran un

patrón similar independientemente de la lateralidad del bateador, mientras que los zurdos ajustan su lanzamiento según enfrenten a un bateador diestro o zurdo. El desplazamiento horizontal de los lanzamientos es distintivo para cada tipo de enfrentamiento, y el desplazamiento vertical presenta diferencias significativas entre los enfrentamientos entre lanzadores zurdos y bateadores diestros y el resto. En resumen, estos resultados indican que existen diferencias significativas en las características de los lanzamientos según el tipo de enfrentamiento lanzador-bateador.

Cuadro 2.7: Medias y códigos de significacción correspondientes a la prueba de de rangos múltiples de Nemenyi sobre la igualdad en distribución de las variables de lanzamientos entre los distintos tipos de enfrentamiento.

Grupos	**Medias**	**Grupos**		
		DD	DI	II
Velocidad				
DD	90,274			
DI	90,394	*		
II	89,044	***	***	
ID	88,682	***	***	***
Localización horizontal				
DD	-0,060			
DI	-0,075			
II	0,022	***	***	
ID	-0,077			***
Localización vertical				
DD	2,400			
DI	2,387			
II	2,413	***	***	
ID	2,362	***		***
Movimiento horizontal				
DD	-1,790			
DI	1,800	***		
II	-2,039	***	***	
ID	2,069	***	***	***
Movimiento vertical				
DD	3,350			
DI	3,352			
II	3,339			
ID	3,437	***	***	***

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *Código de significancia estadísticas:* $\alpha < 0,001 = ***$, $\alpha < 0,01 = **$, $\alpha < 0,05 = *$. *La variable "tipo de enfrentamiento" contiene las categorias; DD, DI, II y ID, según la combinación de brazo de lanzar del lanzador y lado de bateo del bateador (D: diestro, I: zurdo).*

Adicionalmente, se aplicó el mismo procedimiento anterior para determinar si la calidad de los contactos de los bateadores varía según el tipo de enfrentamiento. Específicamente, se compararon la velocidad y el ángulo de salida de la pelota en las diferentes combinaciones de enfrentamientos. La prueba de Kruskal-Wallis (Anexo 4) reveló que las distribuciones de la velocidad y el ángulo de salida de los batazos son estadísticamente distintas entre al menos dos combinaciones de tipos de enfrentamientos. A su vez, la prueba de Nemenyi (Cuadro 2.8) indicó que, en el caso de la velocidad de salida, los enfrentamientos entre lanzadores y bateadores de distintas lateralidades comparten una distribución similar; no obstante, en las demás combinaciones existen diferencias en la distribución. Por otra parte, la distribución del ángulo de salida de los batazos es estadísticamente distinta para todas las combinaciones de enfrentamientos posibles.

Cuadro 2.8: Medias y códigos de significacción correspondientes a la prueba de de rangos múltiples de Nemenyi sobre la igualdad en distribución de "velocidad de salida" y "ángulo de salida" entre los distintos tipos de enfrentamiento.

Grupos	**Medias**	**Grupos**		
		DD	DI	II
Velocidad de salida				
DD	88,436			
DI	89,220	***		
II	86,748	***	***	
ID	88,990	***		***
Ángulo de salida				
DD	12,819			
DI	14,751	***		
II	9,922	***	***	
ID	13,968	**	**	***

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *Código de significancia estadísticas:* $\alpha < 0,001 = ***$, $\alpha < 0,01 = **$, $\alpha < 0,05 = *$. *La variable "tipo de enfrentamiento" contiene las categorias; DD, DI, II y ID, según la combinación de brazo de lanzar del lanzador y lado de bateo del bateador (D: distro, I: zurdo).*

En resumen, el análisis descriptivo ha revelado una relación estadística significativa entre las características de los lanzamientos y los resultados de los batazos. Sin embargo, estos hallazgos preliminares no son suficientes para comprender las interacciones complejas entre las diferentes variables y cómo estas influyen en el resultado final. Además, es fundamental considerar el contexto específico de cada enfrentamiento entre bateador y lanzador. En la siguiente sección, se detallan técnicas estadísticas más avanzadas que permiten profundizar en estas cuestiones.

Capítulo 3

Metodología

El presente capítulo detalla la metodología empleada para analizar la relación entre las características de los lanzamientos y los resultados de los batazos. Con el objetivo de identificar patrones en los lanzamientos asociados a diferentes resultados de bateo, se utiliza un enfoque de modelado flexible basado en técnicas de regresión no paramétrica y semiparamétrica. En primer lugar, se explorarán los modelos de regresión paramétrica y no paramétrica, destacando sus ventajas y limitaciones. Posteriormente, se introducen los modelos de índice único, los cuales permiten modelar de forma flexible la relación entre una variable de respuesta y múltiples variables explicativas. Finalmente, se describirá la estrategia de detección de patrones basada en la estimación de conjuntos de nivel, la cual permitirá identificar las combinaciones de características de los lanzamientos que con mayor probabilidad conducen a un determinado resultado de bateo.

3.1. Análisis de regresión

El análisis de regresión constituye una herramienta estadística fundamental para explorar y cuantificar la relación existente entre una variable de respuesta (Y) y un conjunto de variables predictoras o explicativas $(X_1, X_2, \ldots, X_d)$. Considerando una muestra aleatoria de tamaño n, representada por $\{(Y_i, X_{i1}, X_{i2}, \ldots, X_{id})\}_{i=1}^{n}$,el modelo de regresión se puede expresar de forma general como:

$$Y_i = m(X_{i1}, \ldots, X_{id}) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

donde $m(X_{i1}, \ldots, X_{id}) = E(Y_i, \mid X_{i1}, \ldots, X_{id})$ representa la función de regresión que describe la relación entre la variable respuesta y las predictoras. El término de error aleatorio, ϵ_i, captura la variabilidad no explicada por el modelo, dependiendo del tipo de modelo empleado este término de error debe cumplir determinadas restricciones.

La estimación de la función de regresión $m(X_{i1}, \ldots, X_{id})$ a partir de los datos constituye el objetivo principal del análisis de regresión. En este sentido, existen dos enfoques principales: regresión paramétrica y regresión no paramétrica. Adicionalmente, ambos enfoques se pueden combinar para dar lugar modelos de regresión semiparamétricos. A continuación, se resumen las principales técnicas empleadas bajo estos enfoques.

3.1.1. Regresión paramétrica

Desde el enfoque de la regresión paramétrica, se asume una forma funcional específica para la función m, típicamente lineal o basada en funciones conocidas. Este enfoque es útil cuando se tiene una idea clara de la relación entre la variable dependiente y las variables independientes. Matricialmente,

el modelo de regresión lineal múltiple se puede expresar como (Faraway, 2015, pág. 15):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$, y

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix}.$$

Una suposición es crucial para la validez de los intervalos de confianza y pruebas de hipótesis en el modelo de regresión es asumir que $\boldsymbol{\epsilon} \in N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, siendo σ^2 la varianza del error y $\mathbf{I}_n$ la matriz identidad de orden n.

Una de las principales virtudes del modelo de regresión lineal es la fácil interpretabilidad de los coeficientes de regresión β. En el caso del término de intercepto, $\boldsymbol{\beta}_0$ representa la esperanza de la variable dependiente condicionada a que las covariables sean cero. A su vez, el término β_d correspondiente a cada variable explicativa representa la magnitud del cambio en la esperanza de la variable dependiente ante una variación unitaria en la variable explicativa correspondiente, considerando constante el resto de covariables.

Para la estimación de los coeficientes de regresión en el modelo de regresión lineal múltiple se suele emplear el método de mínimos cuadrados (MMC). Este método consiste en obtener los coeficientes que minimizan la suma de residuos al cuadrado (Montgomery et al. (2012), pag. 72) :

$$\sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \widehat{\boldsymbol{\epsilon}}^T\widehat{\boldsymbol{\epsilon}} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Derivando con respecto a $\boldsymbol{\beta}$ e igualando a cero, se obtiene que el $\widehat{\boldsymbol{\beta}}$ que minimiza los errores al cuadrado satisface:

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}.$$

Por tanto, si $\mathbf{X}^T\mathbf{X}$ es invertible, se obtiene que $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Luego, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. A su vez, la varianza del error se puede estimar como:

$$\widehat{\sigma}^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-p},$$

donde p es el número de coeficientes del modelo, $d+1$ cuando el modelo incluye el intercepto.

Bajo los supuestos estructurales del término de error, se puede realizar inferencias sobre los coeficientes de regresión mediante el estadístico pivote:

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\sigma^2} \sim F_{p,n-p},$$

donde $F_{p,n-p}$ representa la distribución F de Snedecor con p y $n-p$ grados de libertad.

Para realizar inferencia sobre un único coeficiente (β_j), el estadístico pivote sería:

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim T_{n-p},$$

siendo T_{n-p} la distribución T de Student con $n-p$ grados de libertad.

La regresión lineal, a pesar de su simplicidad y facilidad de interpretación, se encuentra limitada por su supuesto de linealidad. Si la relación entre las variables no es lineal, el modelo puede no capturar adecuadamente la complejidad de los datos. Aunque las funciones paramétricas no lineales ofrecen mayor flexibilidad, su uso puede conllevar problemas como la dificultad de estimar con precisión los parámetros y la interpretación de los resultados. Además, tanto la regresión lineal como los modelos paramétricos no lineales se basan en supuestos estrictos sobre la distribución de los errores, los cuales pueden no cumplirse en la práctica. Ante estas limitaciones, los enfoques no paramétricos y semiparamétricos surgen como alternativas atractivas, ya que permiten modelar relaciones más flexibles sin hacer suposiciones tan restrictivas sobre la forma funcional de la relación entre las variables.

3.1.2. Regresión no paramétrica

Mientras que la regresión paramétrica requiere especificar una forma funcional predefinida para la relación entre las variables, la regresión no paramétrica es más relajada en este aspecto. Los métodos no paramétricos no hacen suposiciones fuertes sobre la forma de la función de regresión, sino que la estiman de manera flexible a través de técnicas de suavizado. Algunos ejemplos de métodos no paramétricos son los estimadores tipo núcleo y los splines.

Regresión tipo núcleo

La regresión tipo núcleo se basa en realizar ajustes paramétricos locales, es decir, adaptados al entorno donde se realiza la estima. En un contexto con una única variable explicativa, el estimador de la función de regresión en el punto x se puede aproximar a partir de promedios locales ponderados como:

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n} W_{i,h}(x) Y_i, \quad W_{i,h}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})}$$

donde h es la ventana o parámetro de suavizado que define el entorno de estimación. A su vez, K es una función núcleo, habitualmente una densidad simétrica y centrada en el cero. Esta función asigna pesos decrecientes a cada observación a medida que su distancia con respecto al punto de estimación x aumenta. Este estimador se conoce como *estimador Nadaraya-Watson* o estimador local constante.

En cuanto a las propiedades del estimador local constante, se prueba que, bajo ciertas condiciones de regularidad[1], su error cuadrático medio (MSE) puede aproximarse como (Härdle et al., 2004, pág 92-93):

$$\text{MSE}(\hat{m}_{NW}(x)) = [\text{Sesgo}(\hat{m}_{NW}(x))]^2 + \text{Var}(\hat{m}_{NW}(x)),$$

$$\text{Sesgo}(\hat{m}_{NW}(x))]^2 \approx \frac{h^4}{4}\left[m''(x) + 2\frac{m'(x)f'_X(x)}{f_X(x)}\right]^2 \mu_2(K),$$

$$\text{Var}(\hat{m}_{NW}(x)) \approx \frac{1}{nh}\frac{\sigma^2(x)}{f_X(x)} R(K),$$

donde f_X representa la función de densidad de la variable X, $\mu_2(K) = \int u^2 K(u)\,du$, $R(K) = \int K^2(u)\,du$ y $\sigma^2(x) = Var(Y|X=x)$. A partir de lo anterior, se observa que se trata de un estimador sesgado. La aparición de $m''(x)$ en la expresión del sesgo implica que las zonas de curvatura de la función de regresión serán más difíciles de aproximar. Adicionalmente, se observa que una reducción en el parámetro de suavizado (h) reduciría el sesgo del estimador pero incrementaría la varianza. Por

[1] H1: $m(x)$ es dos veces derivable con derivadas continuas, H2: $\sigma^2(x) = \text{Var}(Y|X=x)$ es continua y positiva, H3: f_x (la función de densidad de X) es derivable con derivada continua, H4: $K(x)$ es una densidad simétrica con varianza finita y cuadrado integrable, H5: $h \equiv h_n$ decrece con n tal que $h_n \to 0$ y $nh_n \to \infty$ cuando $n \to \infty$.

lo tanto, la selección adecuada del parámetro h es crucial.

Existen distintos métodos de selección de la ventana h, uno de los más populares es el de validación cruzada. La idea consiste en encontrar el h que minimiza la función de validación cruzada generalizada que se expresa como:

$$\sum_{i=1}^{n} \left(\frac{Y_i - \hat{m}_{NW}(X_i)}{1 - n^{-1}\text{tr}(S_h)} \right)^2,$$

donde S_h es la matriz de suavizado que contiene los pesos asociados a cada estimación (Wasserman (2006), pag 70).

El estimador de Nadaraya-Watson ajusta constantes o rectas horizontales en cada entorno de x. No obstante, este concepto se puede expandir a otras formas funcionales, lo que da lugar al *estimador polinómico local* de la función de regresión. Si la función de regresión $m(\cdot)$ tiene q derivadas, entonces podría ser aproximada en cada punto x por un polinomio de grado q mediante un desarrollo de Taylor. Este polinomio se puede ajustar minimizando la expresión:

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{q} \beta_j(x)(X_i - x)^j \right)^2 K\left(\frac{X_i - x}{h} \right).$$

De esta forma, se obtendría $\hat{\beta}_j$ como estimador de $\beta_j(x) = \frac{m^{(j)}(x)}{j!}$ con $j = 1, \ldots, p$. En términos prácticos, el vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1 \cdots, \hat{\beta}_p)$ se puede estimar resolviendo un problema de mínimos cuadrados ponderados. Luego, el estimador polinómico local de la función de regresión estaría dado por (Härdle et al. (2004), pag 95):

$$\hat{m}_{PL}(x) = \hat{\beta}_0(x)$$

.

Si el polinomio a ajustar es de grado $p = 0$, el resultado es una constante, es decir, el estimador de Nadaraya-Watson. En el caso de $p = 1$, se obtiene el denominado estimador lineal local.

En términos asintóticos, el estimador lineal local presenta un menor sesgo que el estimador de Nadaraya-Watson, lo que permite utilizar ventanas más grandes que reduzcan la varianza. Adicionalmente, combate mejor el conocido efecto frontera, que consiste en estimaciones sesgadas en los extremos del soporte de X debido a la falta de datos (Wasserman, 2006, pág 80). Por otro lado,la estimación tipo núcleo, además de su dificultad interpretativa, se enfrenta al desastre de la dimensionalidad. Al aumentar el número de variables, se requiere una cantidad exponencialmente mayor de datos para obtener estimaciones confiables. Esto se debe a que la dispersión del espacio de características, dificulta la estimación en regiones poco pobladas. Por esta razón, su uso se limita a problemas con pocas variables explicativas (generalmente 2 o 3).

Regresión por splines

La regresión por spline ofrece una gran flexibilidad para modelar relaciones complejas entre variables. En el caso univariante, el rango de la variable predictora se divide en intervalos y se ajusta un polinomio en cada uno de los intervalos resultantes. Estos polinomios se unen en los puntos de corte, llamados nodos, de forma suave, asegurando la continuidad de la función.

Existen distintas formas de aproximar la función de regresión mediante splines. Una de ellas son los *splines de suavización*, que se basan en la minimización de la suma de residuos al cuadrado penalizados, es decir:

$$M(\lambda) = \sum_{i} (Y_i - \hat{m}_n(x_i))^2 + \lambda J(r),$$

donde $J(r)$ es una penalización por curvatura y el parámetro λ controla el grado de suavidad del ajuste (Wasserman, 2006, pág. 81). Desde su introducción (Reinsch (1967)), la integral del cuadrado de la segunda derivada de la función ajustada se ha convertido en una penalización de suavidad comúnmente utilizada, es decir, $J(r) = \int (r''(x))^2 dx$. Adicionalmente, se suele considerar que la función $\hat{m}_n$ que minimiza $M(\lambda)$ es un spline cúbico natural.

Dado un intervalo $[a, b]$ tal que $a < t_1, t_2, \ldots, t_k < b$, un spline cúbico es una función de la forma:

$$g(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i, \quad \text{para } t \in [t_i, t_{i+1}],$$

con $i \in \{0, \ldots, n\}$, $t_0 = a$ y $t_{n+1} = b$. Para que sea continuo, este spline cúbico debe tener hasta la segunda derivada continua en los nodos. Adicionalmente, para garantizar la unicidad de sus coeficientes, se exige que sea natural, es decir, que su segunda y tercera derivada se anulen en los subintervalos extremos ($d_0 = c_0 = d_n = c_n = 0$). Bajo este contexto, la forma explícita de $m(x)$ estaría dada por:

$$m(x) = \sum_{j=1}^{k+4} \beta_j B_j(x),$$

donde las funciones $\{B_1, \ldots, B_{k+4}\}$ forman una base para el conjunto de splines cúbicos en los nodos $t_1 < t_2 < \ldots < t_k$ contenidos en $[a, b]$. En el caso de la *base de potencia truncada* [2], $B_1(x) = 1$, $B_2(x) = x$, $B_3(x) = x^2$, $B_4(x) = x^3$, $B_j(x) = (x - t_{j-4})_+^3$. En tal sentido, el problema de minimización de los errores al cuadrado penalizados se podría replantear como:

$$(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta},$$

donde $B_{ij} = B_j(X_i)$ y $\boldsymbol{\Omega}$ es una matriz con componente (j,k) dada por $\int B_j''(x) B_k''(x) dx$. Por lo tanto, la búsqueda del óptimo se realiza en una combinación lineal de elementos de una base de splines cúbicos naturales con nodos en cada observación X_i. Derivando, igualando a cero y despejando, se obtiene que:

$$\hat{\beta} = (\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}^T\mathbf{Y}.$$

La selección del parámetro de suavizado λ se puede realizar aplicando algún criterio como validación cruzada o validación cruzada generalizada, utilizando como matriz de suavizado $\mathbf{L} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{B}^T$.

Otra forma de aproximar la función de regresión mediante splines son los *splines de regresión*. La diferencia con los spline de suavización esta en que en lugar de utilizar todas las observaciones como nodos y aplicar una penalización por curvatura para no interpolar, con los splines de regresión se utiliza un número reducido de observaciones como nodos. Por tanto, el grado de suavidad de la aproximación está determinada por la cantidad de nodos utilizados. Un spline cubico de regresión se puede expresar como;

$$m(x) = \sum_{k=1}^{L+3+1} \beta_k B_k(x, t),$$

donde B_k es un elemento de la base definida sobre el conjunto de nodos $t = (t_1, \ldots, t_L)$ y evaluada en el punto x. En este caso los coeficientes β_k se puede estimar mediante mínimos cuadrados.

Además de poder ser tratamos mediante técnicas de estimacion paramétricas, los spline de regresión presentan ventajas computacionales importantes con relación a los spline de suavización, dado que no utilizan todas las observaciones como nodos. Además, brinda la posibilidad de escoger los nodos para capturar mejor características específicas en los datos. No obstante, esta flexibilidad también supone

[2] Debido a la inestabilidad numérica de esta base, en la práctica se emplean otras alternativas, siendo la más popular la base de B-splines (citepBoor1978.

inconvenientes, dado que no existe criterio automático para la selección optima del numero de nodos. Algunos criterios utilizados en la selección del número de nodos están basados en la suma de residuos al cuadrado o algún criterio de información.

Una tercera alternativa que combina las dos aproximaciones es la *regresión spline penalizada*. Como lo indica el nombre, en este caso se incluye en el spline de regresión una penalización que depende de la base considerada. En este contexto, los más utilizados son los B-splines (Eilers and Marx, 1996), que utilizan la base *B-spline* con una penalización basada en los cuadrados de diferencias de coeficientes consecutivos $(\beta_{i+1} - \beta_i)^2$.

La extensión de los spline a múltiples dimensiones es posible mediante el uso de *splines de placa delgada* (*thin plate splines*, en inglés) (Duchon, 1977). Sin embargo, la complejidad computacional de estos modelos puede ser elevada, especialmente en altas dimensiones y gran cantidad de observaciones. Una alternativa que reduce significativamente el costo computacional son los *splines de producto tensorial*. Estos splines construyen una base multidimensional a partir del producto tensorial de bases unidimensionales. La penalización de suavidad se puede aplicar de forma global o marginal, lo que permite controlar la complejidad del modelo y evitar el sobreajuste.Si bien estas técnicas adaptan la estimación por splines al contexto multidimensional, la maldición de la dimensionalidad sigue siendo un desafío, especialmente con grandes conjuntos de datos. La dimensión de la base de los splines de producto tensorial crece exponencialmente con la dimensión del vector de covariables, lo que puede limitar su aplicabilidad.

3.1.3. Regresión semiparamétrica

Los modelos semiparamétricos ofrecen una alternativa flexible para modelar relaciones complejas entre variables, combinando la interpretabilidad de los modelos paramétricos con la capacidad de capturar patrones no lineales de los modelos no paramétricos. Al permitir la especificación tanto de componentes lineales como no lineales en el modelo, estos métodos mitigan el problema de la dimensionalidad asociado a los modelos no paramétricos puros, al tiempo que proporcionan una mayor capacidad de ajuste a los datos. En esta sección, exploraremos en detalle dos clases importantes de modelos semiparamétricos: los modelos parcialmente lineales y los modelos de índice único.

Regresión parcialmente lineal

Para mitigar el problema de la dimensionalidad en la estimación no paramétrica, se puede modelar la relación entre la variable respuesta y un vector de covariables como la suma de dos componentes: uno lineal y otro no paramétrico. Dado una variable dependiente Y y dos subconjuntos de covariables $(X_1, X_2, \ldots, X_d)$ y $(T_1, T_2, \ldots, T_l)$ un modelo parcialmente lineal se podría expresar como:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + s(T) + \epsilon,$$

donde (s) es una función suave (estimable mediante técnicas tipo núcleo o spline). Por tanto, el efecto de las variables ($X_1, \ldots, X_d$) sobre la dependiente sería lineal, mientras que el efecto de las variables $(T_1, \ldots, T_l)$ sería no lineal. No obstante, si la dimensión del vector de covariables $\mathbf{T}$ es elevada, se podrían utilizar *modelos aditivos* para suavizar sin caer en el desastre de la dimensionalidad. De esta forma, se suavizaría de forma univariante en el vector de covariables $\mathbf{T}$, esto es:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + s_1(T_1) + \cdots + s_l(T_l) + \epsilon.$$

Para la estimación de un modelo parcialmente lineal como el planteado, se requiere fijar $E(Y) = \beta_0$ y $E(s_j(T_j)) = 0$ para todo $j \in 1, \ldots, l$, restricciones que garantizan la identificación del modelo.

Existen distintos procedimientos de estimación de $\boldsymbol{\beta}$ y s, siendo uno de los más populares el algoritmo *backfitting* propuesto por Breiman and Friedman (1985). Este es un algoritmo iterativo basado en la alternancia entre la estimación paramétrica y no paramétrica.

La idea de los modelos aditivos se puede extender al caso en que los errores ϵ no sigan un modelo gaussiano. En estos casos, se habla de *modelos aditivos generalizados* (GAM). De esta forma, si el soporte de la media de la variable dependiente es $g : S_Y \to R$, el modelo GAM se puede definir como:

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + s_1(T_1) + \cdots + s_l(T_l),$$

donde g es una función de enlace, por ejemplo, en el caso de que Y no tome valores negativos g podría ser una función gama.

Modelos de índice único

Los *modelos de índice único* (single index model, en inglés) son una herramienta adicional que permite modelar no paramétricamente la relación entre una variable dependiente y múltiples variables explicativas, sin las restricciones de dimensionalidad que enfrentan los modelos no paramétricos. Estos modelos se enfocan en identificar la combinación lineal de las variables explicativas que mejor explica la variabilidad de la variable respuesta, sin hacer suposiciones específicas sobre la forma funcional de la relación. Dado un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_d)^T$, en un modelo de índice único, la función de regresión se puede expresar como:

$$m(\mathbf{X}) = E(Y|\mathbf{X}) = g(\boldsymbol{\beta}^T \mathbf{X}),$$

donde $\boldsymbol{\beta}^T \mathbf{X}$ es una combinación lineal con coeficientes ($\boldsymbol{\beta}$) que captura la información relevante en X para predecir Y y g es una función suave univariante que relaciona la combinación lineal de las covariables con la variable dependiente Y, que puede ser estimada con técnicas tipo núcleo o mediante splines.

La estimación de un modelo de índice único implica encontrar los coeficientes $\boldsymbol{\beta}$ y la función de enlace g. Debido a problemas de identificación asociados a la escala y localización del índice, es común fijar un coeficiente de $\boldsymbol{\beta}$ a 1 y excluir el intercepto.

La estimación de los parámetros en un modelo de índice único se puede realizar en dos etapas o de forma iterativa. La estimación en dos etapas consiste en primero estimar los β de la combinación lineal, asumiendo una forma específica para la función de enlace y aplicando técnicas como mínimos cuadrados semiparamétricos (Ichimura, 1993). Luego, se estima no paramétricamente la función de enlace g a partir del índice estimado.

Por otra parte, la estimación iterativa de los parámetros de un modelo de índice único consiste en estimar los coeficientes de la combinación lineal y la función de enlace en un solo paso iterativo. En este caso, se suele utilizar un algoritmo que alterna entre la estimación del índice y la función de enlace hasta alcanzar convergencia (Hardle et al., 1993, Carroll et al., 1997).

3.2. Conjuntos de nivel

Aunque los modelos de regresión son herramientas poderosas para modelar relaciones entre variables, comprender a fondo la naturaleza de estas relaciones puede ser un desafío. Incluso en los modelos paramétricos sencillos, donde los coeficientes proporcionan una interpretación directa del efecto de cada variable explicativa sobre la variable dependiente, esta interpretación se limita a situaciones en las que las demás variables se mantienen constantes. Además, la presencia de interacciones entre variables puede enmascarar o modificar estos efectos individuales, complicando aún más la interpretación del

modelo.

En el caso de los modelos no paramétricos, la ausencia de coeficientes interpretables dificulta aún más la tarea de aislar el efecto de cada variable. Para enfrentar estos inconvenientes, la estimación de los conjuntos de nivel de la función de regresión puede emplearse como una herramienta complementaria que permite una comprensión más intuitiva de las relaciones subyacentes y la extracción de patrones complejos que de otra manera serían difíciles de detectar.

La estimación de conjuntos de nivel es una técnica utilizada para identificar el dominio de una función donde los valores de la función exceden un umbral específico. Dado un vector de variables aleatorias $\mathbf{X} \in R^d$, una función $\phi : R^d \to R$ y una constante c, el conjunto de nivel de ϕ en c se define como:

$$L(c) = \{x \in R^d : \phi(x) \geq c\}.$$

En consecuencia, $L(c)$ define el subconjunto de características $\mathbf{X}$ donde la función f (que puede representar una densidad, distribución, regresión, etc.) excede el umbral c. La literatura propone diversos métodos para estimar estos conjuntos de nivel, los cuales se clasifican generalmente en dos categorías: directos e indirectos.

La determinación directa de los conjuntos de nivel de una función de regresión elude la necesidad de estimar previamente dicha función. Bajo esta perspectiva, diversos métodos proponen abordar la estimación de $L(c)$ como un problema de clasificación binaria. A modo de ejemplo, Willett and Nowak (2007) introducen un estimador basado en particiones diádicas recursivas, mientras que Scott and Davenport (2007) emplean una estrategia de clasificación sensible al coste. Alternativamente, la estimación directa puede plantearse como un problema de optimización. Existe una amplia literatura que explora esta perspectiva desde el enfoque bayesiano, donde la función de regresión se modela como una variable aleatoria y se aproxima $L(c)$ mediante muestreo secuencial guiado por una función de adquisición que cuantifica la incertidumbre sobre el conjunto de nivel (Bryan et al., 2005; Bogunovic et al., 2016; Shekhar and Javidi, 2019; Ha et al., 2021). Otro enfoque de optimización, propuesto por Cavalier (1997) y Polonik and Wang (2005), se basa en la maximización del exceso de masa para estimar $L(c)$.

Por otro lado, la estimación indirecta o plug-in de los conjuntos de nivel implica reemplazar la función desconocida f por un estimador adecuado $\hat{f}$ obtenido a partir de los datos. En el caso de la regresión, $\hat{f} = \hat{m}(x)$, donde $\hat{m}(x)$ es cualquier estimador de la función de regresión. Habitualmente se emplea un estimador no paramétrico de la función de regresión, las propiedades teóricas $\hat{L}(c)$ han sido estudiadas en profundidad en el contexto de la regresión tipo núcleo por Laloë and Servien (2013) y Dau et al. (2020). Aunque este método es más sencillo, también es más sensible al problema de la dimensionalidad que los métodos directos debido a las limitaciones inherentes a los estimadores no paramétricos en altas dimensiones. Asimismo, la identificación de los elementos del conjunto de nivel $L(c)$ en espacios de covariables de dimensión superior a dos podría suponer un desafío computacional significativo.

3.3. Estrategia de detección de patrones en lanzamientos

Los objetivos del estudio se centran en identificar patrones en las características de los lanzamientos que influyen en resultados de bateo determinados. Específicamente, los resultados de bateo se definen en términos de la velocidad o ángulo vertical de salida de la pelota al ser bateada, mientras que los lanzamientos se caracterizan por su velocidad, localización (vertical y horizontal) y desplazamiento o rompimiento (vertical y horizontal).

La estimación de la relación entre las características de los lanzamientos y las variables dependientes se aborda individualmente para cada variable dependiente desde la perspectiva de los modelos de regresión de spline aditiva. La adopción de modelos spline aditivos responde no solo a la necesidad de lidiar con el problema de la dimensión, sino también al interés en centrar el análisis en los efectos principales de las características de los lanzamientos sobre las variables dependientes analizadas, lo cual podría ser más complicado si se permite la interacción entre las variables explicativas. Adicionalmente, la comparación mediante análisis de varianza (ANOVA) de los modelos spline aditivos y modelos de índice único indica que los primeros explican una mayor proporción de la variabilidad en la variable dependiente.

La velocidad de lanzamiento por sí sola no representa una característica sobre la cual los lanzadores tengan mucho margen de control con resultados beneficiosos, dadas las limitaciones físicas de cada lanzador. Por lo tanto, la búsqueda de patrones se centra en la localización y el desplazamiento de los lanzamientos. Para ello, se estiman los modelos:

$$g(E(Y^j)) = \beta_0 + s_1(\text{plate}_x) + s_2(\text{plate}_z) + s_3(\text{velocidad}_l) + \epsilon, \quad j = 1, 2,$$

$$E(Y^j) = \beta_0 + s_1(\text{mov}_x) + s_2(\text{mov}_z) + s_3(\text{velocidad}_l) + \epsilon, \quad j = 1, 2,$$

donde Y^1 y Y^2 representan la velocidad y ángulo vertical de salida de bateo respectivamente, siendo la función de enlace g la identidad en el primer caso y una gamma en el segundo. En cuanto a las variables explicativas, plate_x y plate_z representan las coordenadas verticales y horizontales de los lanzamientos, mientras que mov_x y mov_z representan el desplazamiento horizontal y vertical de los lanzamientos, respectivamente. Finalmente, velocidad_l es la velocidad de los lanzamiento.

A partir de la estimación de los modelos especificados, se extraen patrones de localización o desplazamiento de lanzamientos fijando la velocidad de lanzamiento en el promedio de la muestra y construyendo los contornos de los conjuntos de nivel de $L(c) = \{\mathbf{x} \in R^3 : E(Y^j) \geq c\}$ para distintos puntos de corte c ($\mathbf{x} \in \{\text{plate}_x, \text{plate}_z, \text{velocidad}_l\}$ o $\mathbf{x} \in \{\text{mov}_x, \text{mov}_z, \text{velocidad}_l\}$ segun corresponda). Lo anterior da como resultado gráficos de curvas de nivel que indican patrones de localización (horizontal o vertical) o desplazamiento (horizontal o vertical) de lanzamientos que dan lugar a determinados valores esperados en velocidad o ángulo vertical de salida de bateo.

No obstante, tanto los tipos de enfrentamiento como cada tipo de lanzamiento exhiben realidades distintas tanto en términos de características de lanzamientos como de resultados de bateo. Por esta razón, los datos se agrupan por lateralidad de los enfrentamientos (según lanzador o bateador compartan o no lateralidad) y tipo de lanzamiento (recta de 4 costuras, sinker, slider, cambio, cutter, sweeper y curva), y se estima un modelo distinto en cada grupo.

Capítulo 4

Resultados empíricos

En este capítulo se presentan los resultados del análisis realizado para identificar los patrones de ubicación y desplazamiento de los lanzamientos que influyen en la velocidad y el ángulo de salida de la pelota al ser bateada. Mediante la visualización de curvas de nivel, se identifican las zonas y los tipos de movimiento que maximizan o minimizan la probabilidad de obtener determinados resultados. En estos gráficos, las zonas más oscuras representan valores más bajos en la esperanza de la variable dependiente, indicando áreas donde los lanzamientos son menos efectivos para el bateador. Por el contrario, las zonas más claras indican valores más altos, señalando áreas donde los lanzamientos son más propensos a ser conectados con mayor fuerza o ángulo. Los gráficos de curvas de nivel se presentan para cada tipo de enfrentamiento y tipo de lanzamiento, proporcionando una visión detallada de cómo la ubicación y el desplazamiento de los lanzamientos afectan el rendimiento de los bateadores.

4.1. Análisis de la velocidad de bateo

En este apartado se aborda el análisis de cómo la ubicación de los lanzamientos influye en la capacidad de los bateadores para generar potencia. A través de la visualización de los gráficos de curvas de nivel, se identificarán las zonas más vulnerables para los bateadores al enfrentar diferentes tipos de lanzamientos, tanto en enfrentamientos entre jugadores de la misma lateralidad como en aquellos donde la lateralidad es diferente

4.1.1. Patrones en la localización de los lanzamientos

A continuación, se explora cómo la ubicación de los lanzamientos influye en la velocidad con la que la pelota es bateada, tanto en enfrentamientos entre bateadores y lanzadores de la misma lateralidad como en aquellos donde la lateralidad es diferente. A través del análisis de datos y la visualización de curvas de nivel, se identificarán las zonas más vulnerables para los bateadores al enfrentar diferentes tipos de lanzamientos en cada una de estas situaciones.

Enfrentamientos entre lanzadores y bateadores de igual lateralidad

En los enfrentamientos entre lanzadores y bateadores de la misma lateralidad, los lanzamientos localizados en las esquinas inferiores de la zona de bateo suelen generar los contactos de bateo más débiles. Sin embargo, un análisis más profundo revela patrones específicos para cada tipo de lanzamiento.

Para la recta de 4 costuras, se observa que la zona de menor velocidad de salida se concentra en la parte inferior interna de la zona de strike, formando una especie de "Lïnvertida. Esta tendencia se alinea con la estrategia común de mantener las rentas bajas y/o adentro para dificultar el swing del

bateador, especialmente considerando la alta velocidad de este lanzamiento, lo cual dificulta que el bateador genere un swing fluido y potente.

Por otra parte, los resultados indican que el sinker, slider y cambio generan las velocidades de salida más bajas cuando se ubican en la zona interna. El movimiento lateral del sinker y slider, y el desplazamiento vertical abrupto del cambio, hacen que el bateador conecte no conecte con la parte central del bate o con un swing adelantado, respectivamente. Esto resulta en contactos menos sólidos y velocidades de salida reducidas.

En lo que respecta al cutter y el sweeper, su zona de localización que inducen velocidades de contactos bajas se encuentran en las equinas inferiores de la zona de bateo. El cutter y el sweeper, el igual que la slider y el sinker que se caracterizan por el desplazamiento vertical, no obstante, se lanzan a menor velocidad. Al ser lanzamientos más lentos, los bateadores tienen más tiempo para ajustar su swing si el lanzamiento se localiza en la parte alta, siendo más probable que conecten con mayor fortaleza. La curva, por su parte, induce velocidades de contacto más bajas cuando se localiza en la esquina externa de la zona de bateo. La trayectoria y la velocidad relativamente baja de este lanzamiento podrían inducir a los bateadores a intentar "tirar" de la pelota, resultando en conexiones débiles y poca potencia.

Enfrentamientos entre lanzadores y bateadores de diferente lateralidad

En los enfrentamientos entre lanzadores y bateadores de distinta lateralidad (diestros contra zurdos o viceversa), los patrones de localización que inducen velocidades de contacto bajas en el slider son sorprendentemente similares a los observados en enfrentamientos entre bateadores y lanzadores de la misma lateralidad. A pesar de la diferente perspectiva del bateador, los sliders en la zona interna siguen siendo los que se conectan con menor velocidad. El sinker, por su parte, muestra un comportamiento similar, pero con una peculiaridad: los lanzamientos en la esquina inferior externa también reducen significativamente la velocidad de bateo, un efecto más pronunciado en estos enfrentamientos.

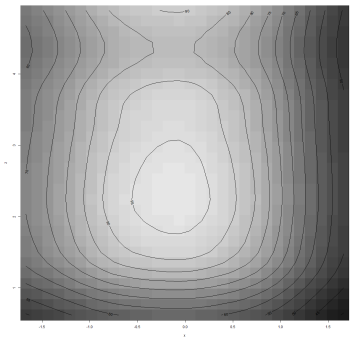
En lo que respecta a la recta de 4 costuras, los lanzamientos en la parte interna de la zona de bateo son claramente los que inducen la menor velocidad de bateo, lo mismo ocurre con el cutter y el sweeper. A diferencia de lo observado en los enfrentamientos entre lanzadores y bateadores de distintas lateralidades, la esquina inferior externa de la zona de bateo no es una zona segura para los lanzadores en términos de velocidad de bateo esperada. Este resultado tiene bastante sentido, ya que en los enfrentamientos entre lanzadores y bateadores de distintas lateralidades, el desplazamiento horizontal de los lanzamientos se produce desde fuera hacia el bateador. Por lo tanto, mientras más afuera de la zona de bateo se localice el lanzamiento, mayor coincidencia existe con la trayectoria del bate durante el swing, produciendo contactos más sólidos.

La combinación del movimiento de los lanzamientos y la trayectoria del swing del bateador explica las zonas de menor velocidad de salida en el cambio y la curva. El cambio induce velocidades más bajas cuando se ubica en la esquina inferior externa. Al carecer de un desplazamiento horizontal significativo, en esta zona los cambios pueden producir swings desbalanceados o fuera de tiempo sin caer en el peligro de coincidencia entre la trayectoria del bate y el desplazamiento horizontal del lanzamiento. Por su parte, las curvas, con su pronunciado movimiento vertical y horizontal, generan velocidades más bajas en ambas esquinas inferiores, siendo más notable en la esquina interna, dado que la trayectoria entre bate y pelota coinciden menos.

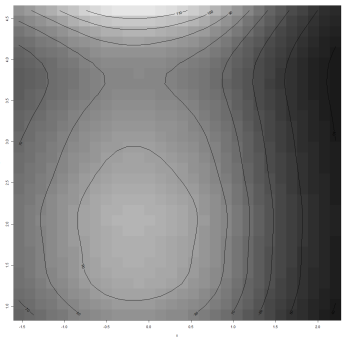
4.1.2. Patrones en el desplazamiento de los lanzamientos

En general, un mayor movimiento se asocia con una menor velocidad de salida. Sin embargo, analizar la dirección de este movimiento, ya sea vertical u horizontal, nos permite identificar patrones

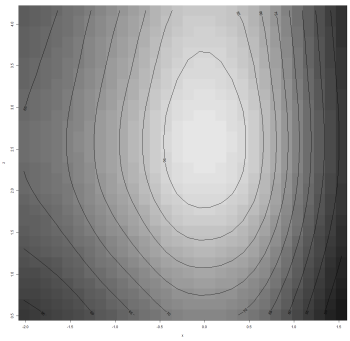
Gráfico 4.1: Zonas de mayor y menor velocidad de bateo esperada según la localización del lanzamiento en enfrentamientos entre lanzadores y bateadores de igual lateralidad



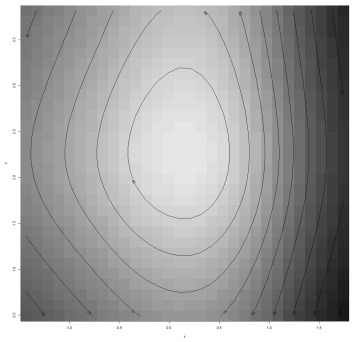
(a) Recta de 4 costuras



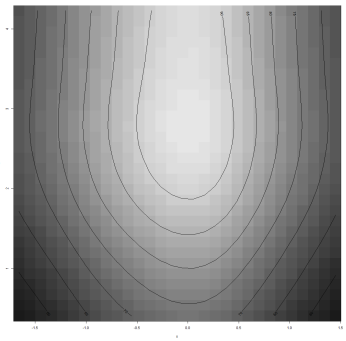
(b) Sinker



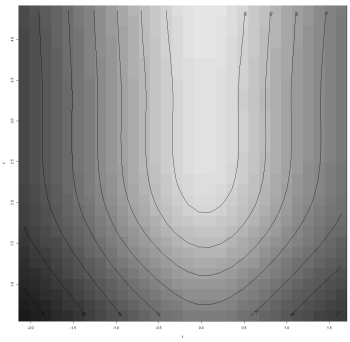
(c) Slider



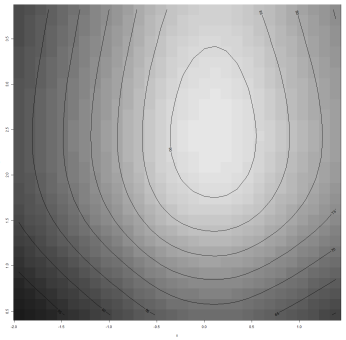
(d) Cambio



(e) Cutter



(f) Sweeper

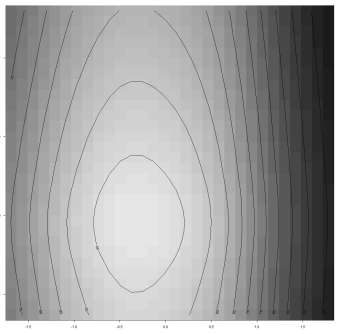


(g) Curva

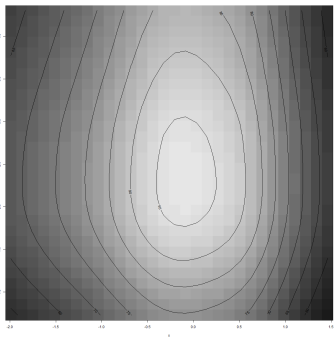
Notas: *Los ejes X e Y representan las coordenadas horizontales y verticales respectivamente. Valores positivos en X indican una distancia desde el centro del plato de bateo hacia el bateador. Los valores en Y representan la altura sobre el suelo. La velocidad se lanzamiento se fija en el promedio de la muestra.*

más específicos. Por ejemplo, podemos determinar cuál de las componentes del movimiento de un lanzamiento, como una curva, tiene un mayor impacto en reducir la velocidad de salida.

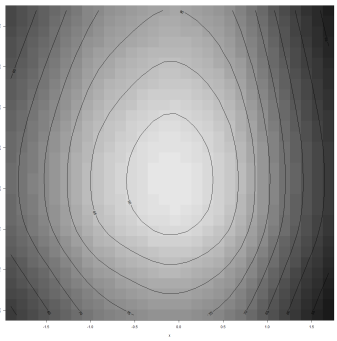
Gráfico 4.2: Zonas de mayor y menor velocidad de bateo esperada según la localización del lanzamiento en enfrentamientos entre lanzadores y bateadores de distinta lateralidad



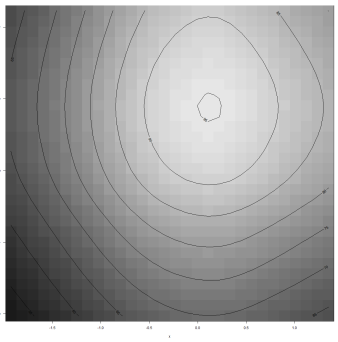
(a) Recta de 4 costuras



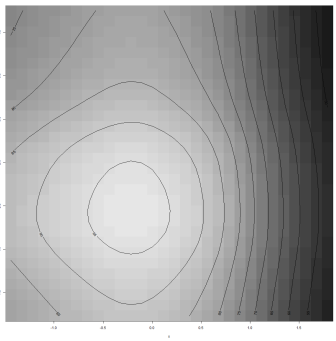
(b) Sinker



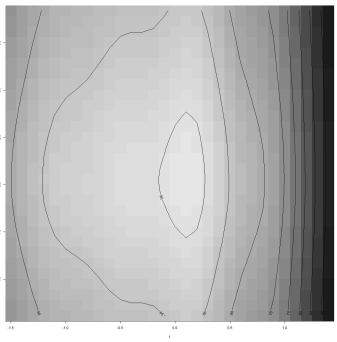
(c) Slider



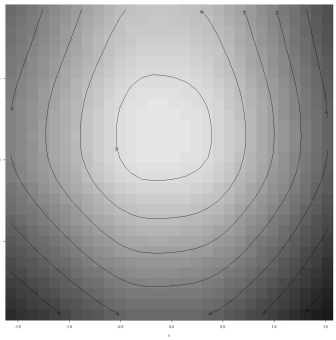
(d) Cambio



(e) Cutter



(f) Sweeper



(g) Curva

Notas: *Los ejes X e Y representan las coordenadas horizontales y verticales respectivamente. Valores positivos en X indican una distancia desde el centro del plato de bateo hacia el bateador. Los valores en Y representan la altura sobre el suelo. La velocidad se lanzamiento se fija en el promedio de la muestra.*

Enfrentamientos entre lanzadores y bateadores de igual lateralidad

En el Gráfico 4.3 se muestran las curvas de nivel estimadas para analizar los patrones de desplazamientos en los enfrentamientos entre lanzadores y bateadores de distintas lateralidades. En cuanto a la recta de 4 costuras, se observa que la amplificación de su movimiento característico hacia la parte externa de la zona de bateo implica una reducción en la velocidad de bateo esperada. A su vez, el

sinker exhibe una zona de baja velocidad esperada cuando su desplazamiento vertical es bajo. Para entender este resultado, es importante recordar que el sinker es un tipo particular de recta, por lo tanto, un mayor desplazamiento vertical solo se alcanza a costa de una reducción de la velocidad, lo que implicaría una menor efectividad del lanzamiento dado que su rango de movimiento horizontal es muy reducido.

El cambio y el slider presentan patrones interesantes en cuanto a la velocidad de salida de la pelota. El cambio, aunque su desplazamiento vertical sea pronunciado, resulta más efectivo cuando además tiene un componente horizontal hacia el exterior de la zona de strike. Por su parte, el slider y la curva muestran una mayor reducción en la velocidad de salida cuando su movimiento combina tanto un desplazamiento vertical como horizontal. El cutter y el sweeper, en cambio, parecen depender más del desplazamiento vertical para reducir la velocidad de salida.

Enfrentamientos entre lanzadores y bateadores de diferente lateralidad

En el caso de los enfrentamientos entre lanzadores y bateadores de distintas lateralidades, el Gráfico 4.4 muestra que la recta de cuatro costuras es más efectiva en la reducción de la velocidad de bateo esperada cuando se desplaza hacia la parte externa de la zona de bateo, un movimiento opuesto al patrón detectado en los enfrentamientos entre jugadores de igual lateralidad.

En cuanto al sinker, tanto el desplazamiento vertical como el horizontal hacia la parte externa de la zona de bateo son relevantes para inducir velocidades de bateo esperadas reducidas. A diferencia de los enfrentamientos entre lanzadores y bateadores de igual lateralidad, en este caso la velocidad del lanzamiento no es tan relevante, siempre y cuando se conserve cierto grado de desplazamiento horizontal. Esto podría deberse a que el sinker, al tener un movimiento más pronunciado, es más difícil de conectar sólidamente cuando se desplaza hacia afuera.

Para el slider, el patrón es similar al encontrado en los enfrentamientos entre lanzadores y bateadores de la misma lateralidad, aunque en este caso el desplazamiento que induce velocidades de contacto bajas es hacia la parte externa de la zona de bateo, debido al cambio de perspectiva del bateador. En cuanto al cambio, su efectividad en la reducción de la velocidad de bateo se magnifica con el desplazamiento vertical y horizontal hacia fuera, típico del lanzamiento. Esto puede explicarse porque el cambio, al parecerse inicialmente a una recta, engaña al bateador, y su desplazamiento adicional hacia afuera dificulta aún más un contacto sólido.

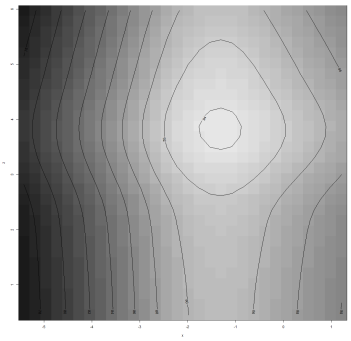
Para el cutter, a diferencia de lo que ocurría en los enfrentamientos entre lanzadores y bateadores de la misma lateralidad, el desplazamiento horizontal hacia el bateador es el principal factor reductor de la velocidad de bateo esperada. Dado el tipo de enfrentamiento, este es el movimiento natural de este tipo de lanzamiento.

En cuanto al sweeper, se observa que en general induce velocidades de bateo relativamente bajas en comparación con otros lanzamientos. No obstante, los impactos positivos de distintas combinaciones de desplazamiento solo se conseguirían con desplazamientos horizontales hacia fuera con respecto al bateador, un movimiento poco frecuente en este tipo de lanzamientos en el contexto del enfrentamiento analizado.

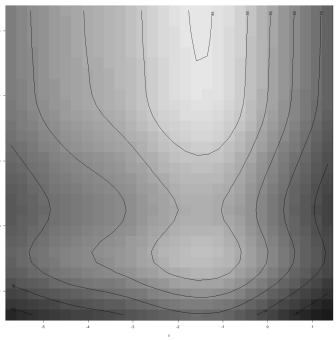
Finalmente, la curva exhibe el mismo comportamiento que en los enfrentamientos entre lanzadores y bateadores de igual lateralidad, siendo el desplazamiento vertical el factor predominante en la reducción de la velocidad de contacto de bateo.

En resumen, se observa que en los enfrentamientos entre lanzadores y bateadores de distintas lateralidades, los lanzamientos que se alejan del bateador son más difíciles de conectar con fuerza. Este

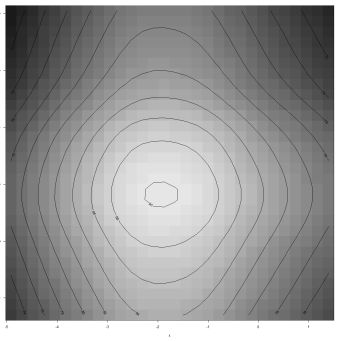
Gráfico 4.3: Zonas de mayor y menor velocidad de bateo esperada según el desplazamiento de los lanzamientos en enfrentamientos entre lanzadores y bateadores de igual lateralidad



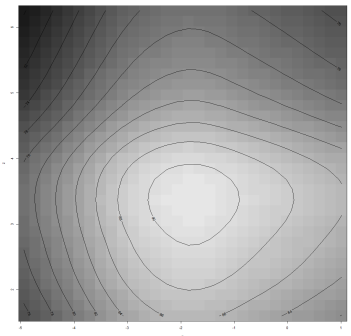
(a) Recta de 4 costuras



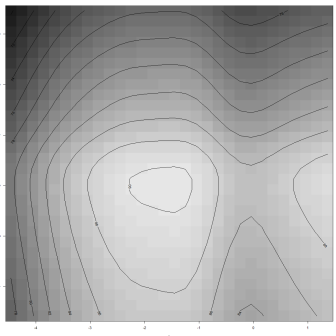
(b) Sinker



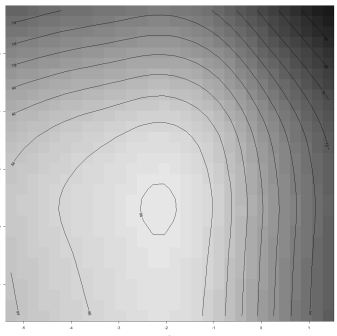
(c) Slider



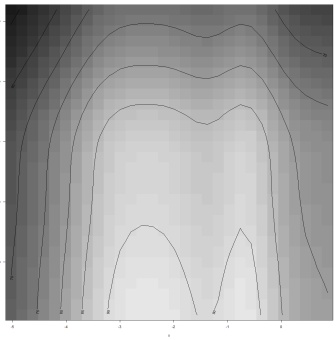
(d) Cambio



(e) Cutter



(f) Sweeper



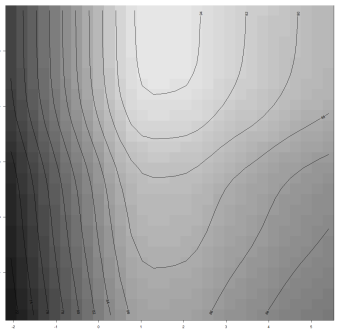
(g) Curva

Notas: *Los ejes X e Y representan las coordenadas horizontales y verticales respectivamente. Valores positivos en X indican una distancia desde el centro del plato de bateo hacia el bateador. Los valores en Y representan la altura sobre el suelo. La velocidad se lanzamiento se fija en el promedio de la muestra.*

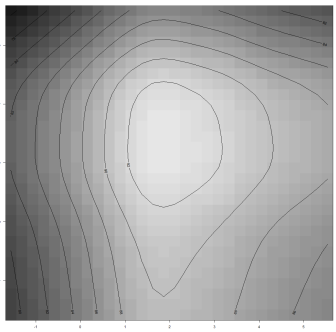
resultado es bastante obvio, dado que en este contexto de enfrentamiento el movimiento predominante natural de los lanzamientos es hacia fuera con respecto al bateador. No obstante, algo que sí aportan estos resultados es que el movimiento horizontal en los lanzamientos es siempre un factor fundamental en la reducción de la velocidad de contacto esperada en bateadores en enfrentamientos entre jugadores de distintas lateralidades, lo que no siempre se cumple en otros tipos de enfrentamientos. Esto puede explicarse por la diferente perspectiva del bateador, que encuentra más difícil ajustar su swing a

lanzamientos que se alejan de su cuerpo, mientras que en otros tipos de enfrentamientos la dificultad de ajuste del swing estará más influenciada por la localización del lanzamiento.

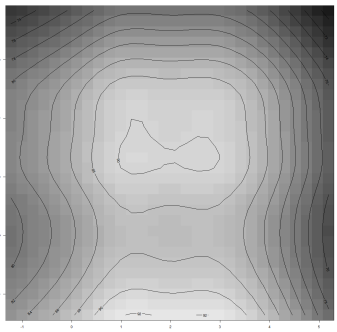
Gráfico 4.4: Zonas de mayor y menor velocidad de bateo esperada según el desplazamiento de los lanzamientos en enfrentamientos entre lanzadores y bateadores de distinta lateralidad



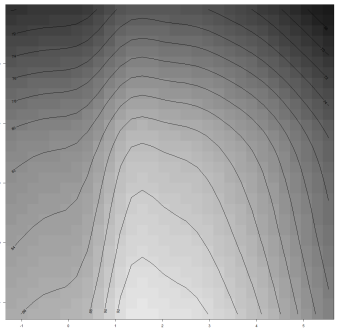
(a) Recta de 4 costuras



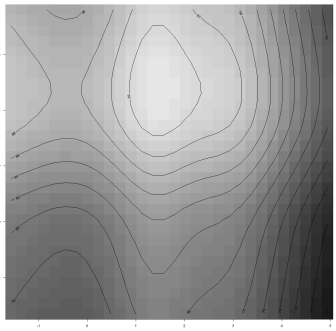
(b) Sinker



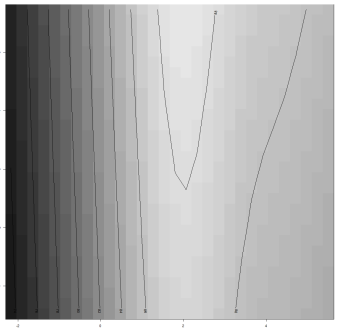
(c) Slider



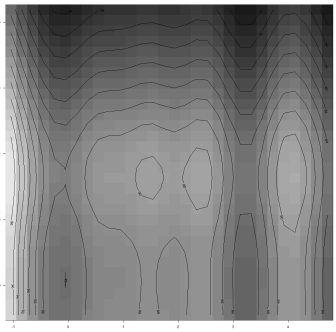
(d) Cambio



(e) Cutter



(f) Sweeper



(g) Curva

Notas: *Los ejes X e Y representan las coordenadas horizontales y verticales respectivamente. Valores positivos en X indican una distancia desde el centro del plato de bateo hacia el bateador. Los valores en Y representan la altura sobre el suelo. La velocidad se lanzamiento se fija en el promedio de la muestra.*

4.2. Análisis del ángulo vertical de bateo

En este apartado exploraremos cómo la ubicación y el movimiento de los lanzamientos influyen en la probabilidad de que la pelota sea bateada por el suelo, tanto en enfrentamientos entre bateadores y lanzadores de la misma lateralidad como en aquellos donde la lateralidad es diferente. A través del análisis de los ángulos de salida de la pelota y la visualización de curvas de nivel, se identificarán las zonas de la zona y los tipos de movimiento que favorecen la generación de batazos rodados.

4.2.1. Patrones en la localización de los lanzamientos

El análisis de los ángulos de salida de la pelota en relación con la ubicación vertical y horizontal de los lanzamientos revela que, en general, los lanzamientos bajos tienden a generar rodados (ángulos de salida inferiores a 10°). Sin embargo, este patrón presenta variaciones significativas dependiendo del tipo de lanzamiento y del enfrentamiento en particular.A continuación, se detallan estas particularidades para enfrentamientos entre lanzadores y bateadores de igual y distinta lateralidad.

Enfrentamientos entre lanzadores y bateadores de igual lateralidad

En el Gráfico 4.5 se presentan las curvas de nivel correspondientes a los ángulos de salida de la pelota en enfrentamientos entre lanzadores y bateadores de la misma lateralidad. En general, los lanzamientos localizados en la esquina inferior externa de la zona de strike tienden a inducir ángulos de salida más bajos, es decir, a generar más rodados. Este patrón se observa de manera consistente en rectas de cuatro costuras, cambios, cutters, sweepers y curvas. La combinación de la ubicación baja y externa parece dificultar que el bateador eleve la pelota, favoreciendo un contacto más débil y un ángulo de salida más cerrado.

Sin embargo, el sinker presenta un comportamiento ligeramente diferente. Los ángulos de salida más bajos se obtienen cuando el sinker se localiza en la parte baja y/o en la parte interior de la zona de bateo. Este comportamiento se explica por el movimiento natural del sinker, que se hunde hacia abajo y hacia adentro, dificultando que el bateador eleve la pelota.

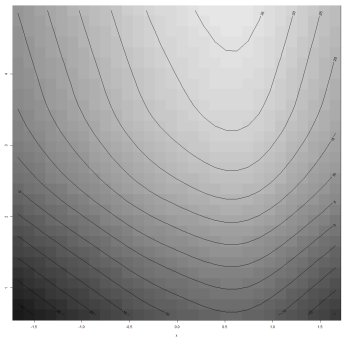
La slider, por su parte, exhibe dos zonas destacadas de influencia en la generación de batazos rodados: la esquina inferior interna y la externa. Esto se debe a la combinación de su movimiento lateral y su caída, que dificulta que el bateador haga un buen contacto con la pelota y la eleve con fuerza.

Enfrentamientos entre lanzadores y bateadores de diferente lateralidad

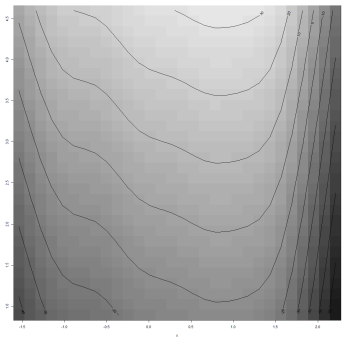
En los enfrentamientos entre lanzadores y bateadores de distinta lateralidad, el análisis de los ángulos de salida de la pelota revela un patrón consistente: los lanzamientos localizados en la esquina inferior externa de la zona de bateo tienden a generar una mayor proporción de rodados. Este resultado coincide con los hallazgos obtenidos en enfrentamientos entre bateadores y lanzadores de la misma lateralidad, donde también se observó una relación entre la ubicación baja y externa del lanzamiento y una disminución en el ángulo de salida.

Sin embargo, existen algunas diferencias notables entre ambos tipos de enfrentamientos. Mientras que en los enfrentamientos entre bateadores y lanzadores de la misma lateralidad los patrones de localización óptima para inducir roletazos variaban ligeramente según el tipo de lanzamiento, en los enfrentamientos entre bateadores y lanzadores de distinta lateralidad, la esquina inferior externa se consolida como la zona más efectiva para generar roletazos, independientemente del tipo de lanzamiento. Esta tendencia general a generar más roletazos en los enfrentamientos entre bateadores y lanzadores de distinta lateralidad puede explicarse por la mayor dificultad que supone hacer contacto sólido con

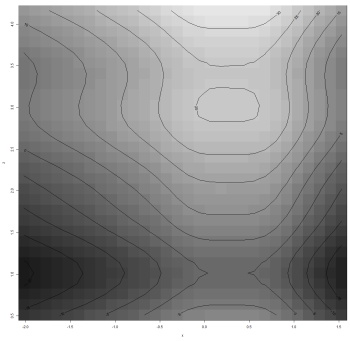
Gráfico 4.5: Zonas de mayor y menor angulo vertical de bateo esperado según la localización de los lanzamientos en enfrentamientos entre lanzadores y bateadores de igual lateralidad



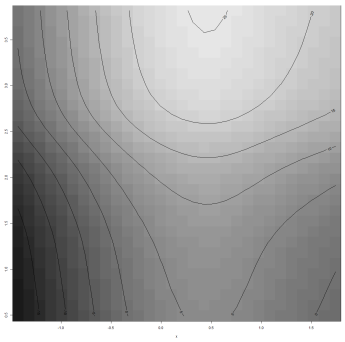
(a) Recta de 4 costuras



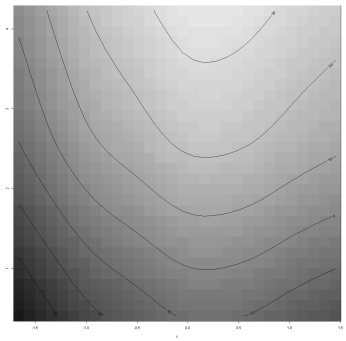
(b) Sinker



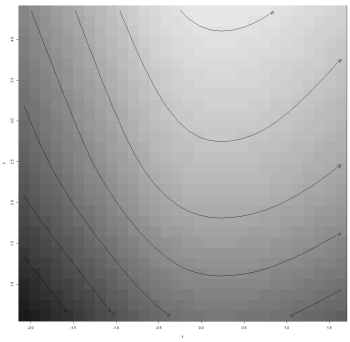
(c) Slider



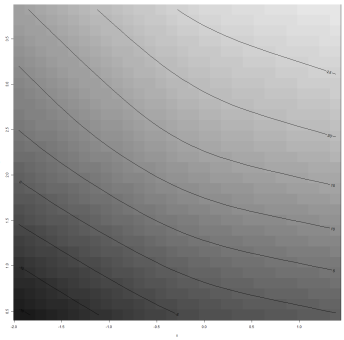
(d) Cambio



(e) Cutter



(f) Sweeper



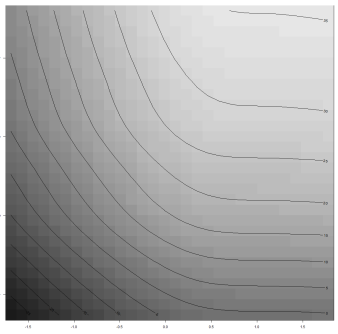
(g) Curva

Notas: *Los ejes X e Y representan respectivamente el desplazamiento (pies) horizontales y verticales respectivamente de los lanzamientos. La velocidad se lanzamiento se fija en el promedio de la muestra para cada tipo de lanzamiento.*

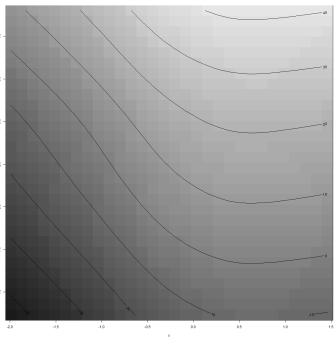
la pelota en una zona más alejada del cuerpo.

Adicionalmente, se observa que en los enfrentamientos entre lanzadores y bateadores de distinta lateralidad, los ángulos verticales de bateo esperados son generalmente inferiores a los observados en otros tipos de enfrentamientos. Es decir, en este contexto resulta más difícil para el lanzador inducir batazos rodados.

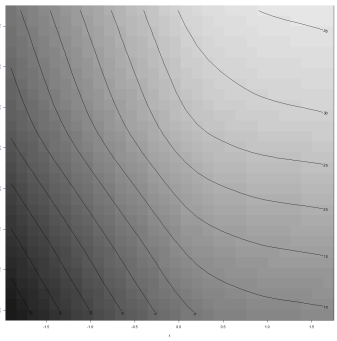
Gráfico 4.6: Zonas de mayor y menor angulo vertical de bateo esperado según la localización de los lanzamientos en enfrentamientos entre lanzadores y bateadores de distinta lateralidad



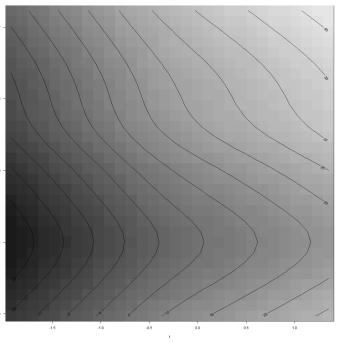
(a) Recta de 4 costuras



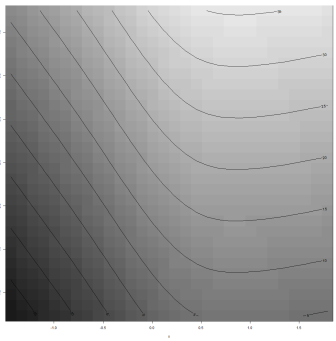
(b) Sinker



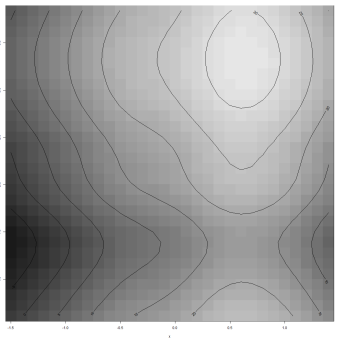
(c) Slider



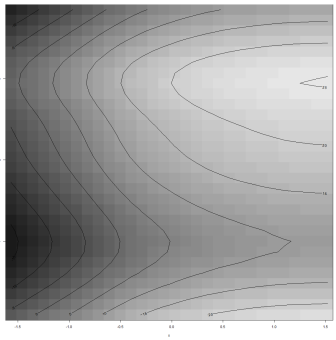
(d) Cambio



(e) Cutter



(f) Sweeper



(g) Curva

Notas: *Los ejes X e Y representan respectivamente el desplazamiento (pies) horizontales y verticales respectivamente de los lanzamientos. La velocidad se lanzamiento se fija en el promedio de la muestra para cada tipo de lanzamiento.*

4.2.2. Patrones en el desplazamiento de los lanzamientos

A continuación, se analiza la influencia del desplazamiento de los lanzamientos en el ángulo vertical de salida de los batazos. Los resultados obtenidos revelan que, en general, el movimiento vertical de los lanzamientos es el factor determinante en la generación de batazos rodados. Sin embargo, al igual que en el caso de los patrones relacionados con la localización, el análisis detallado evidencia

particularidades interesantes.

Enfrentamientos entre lanzadores y bateadores de igual lateralidad

En los enfrentamientos entre lanzadores y bateadores de la misma lateralidad, el Gráfico 4.7 revela un patrón predominante: las curvas de nivel que representan los ángulos de salida esperados tienden a ser horizontales, indicando que el desplazamiento vertical del lanzamiento ejerce una mayor influencia en el ángulo de salida que el desplazamiento horizontal.

En general, un mayor movimiento vertical descendente del lanzamiento se asocia con ángulos de salida más bajos, es decir, con una mayor probabilidad de generar rodados. Este resultado es intuitivo, ya que un lanzamiento bajo tiende a inducir un contacto más elevado en la pelota, favoreciendo un ángulo de salida más cerrado.

Sin embargo, el análisis detallado revela algunas particularidades. En el caso de las curvas, se observa una sinergia entre el movimiento vertical descendente y el desplazamiento horizontal hacia el exterior de la zona de strike, lo que resulta en ángulos de salida aún más bajos.

Por otro lado, los patrones observados para el sinker y el cambio son menos claros, con variaciones en los ángulos de salida esperados para diferentes combinaciones de desplazamiento vertical y horizontal. Esta aparente ambigüedad podría deberse a la limitada variabilidad en los ángulos de salida predichos para estos lanzamientos, especialmente en el caso del sinker, donde se predicen principalmente roletazos para el rango de desplazamientos analizado.

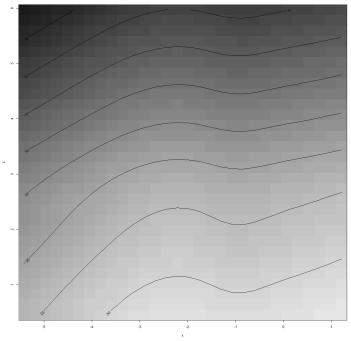
Enfrentamientos entre lanzadores y bateadores de diferente lateralidad

En los enfrentamientos entre lanzadores y bateadores de distinta lateralidad, el Gráfico 4.8 confirma la tendencia observada en enfrentamientos entre bateadores y lanzadores de la misma lateralidad: el desplazamiento vertical del lanzamiento es el principal factor que influye en la disminución del ángulo vertical de salida de la pelota.

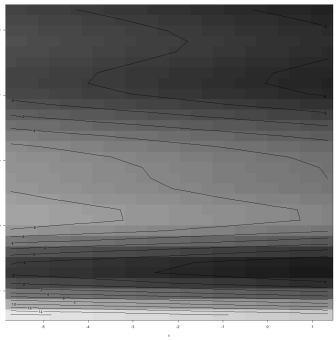
Sin embargo, se identifican algunas particularidades en este tipo de enfrentamientos. Por ejemplo, en el caso de los sliders y cambios, la combinación de un movimiento vertical descendente y un desplazamiento hacia el interior de la zona de bateo resulta en ángulos de salida aún más cerrados. De manera similar, las curvas, sweepers y cutters son más efectivos en generar roletazos cuando se combinan un movimiento vertical pronunciado y un desplazamiento hacia el exterior de la zona de bateo.

En resumen, independientemente del tipo de enfrentamiento, el movimiento vertical del lanzamiento es el factor más relevante para inducir ángulos de salida bajos. El sinker, en particular, se destaca por su consistencia en generar rodados, debido a su movimiento abrupto natural hacia abajo.

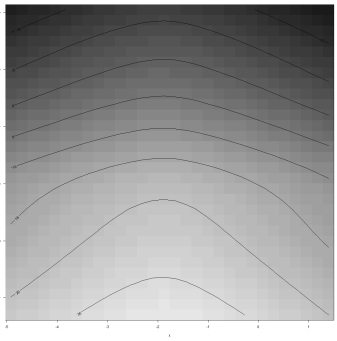
Gráfico 4.7: Zonas de mayor y menor angulo vertical de bateo esperado según el desplazamiento de los lanzamientos en enfrentamientos entre lanzadores y bateadores de igual lateralidad



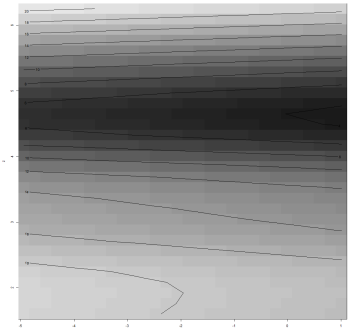
(a) Recta de 4 costuras



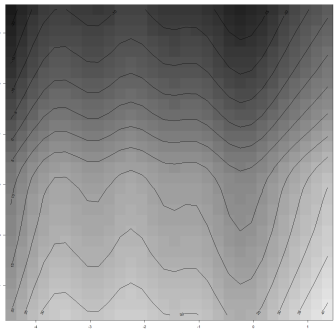
(b) Sinker



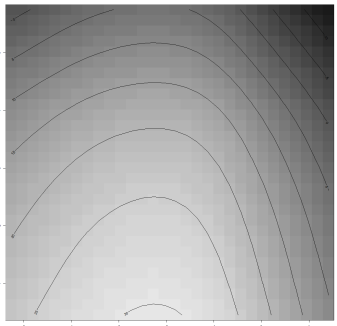
(c) Slider



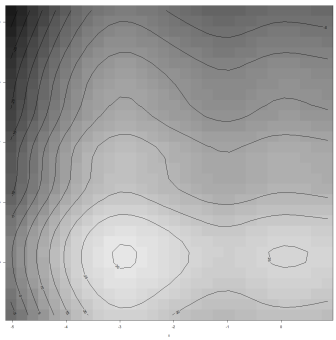
(d) Cambio



(e) Cutter



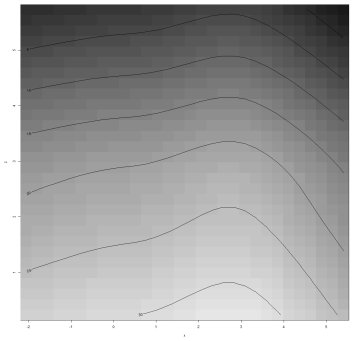
(f) Sweeper



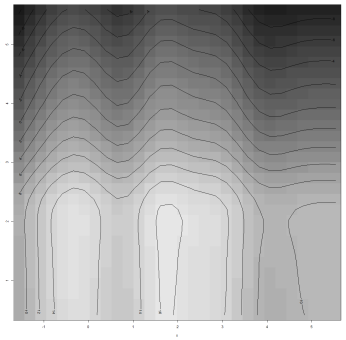
(g) Curva

Notas: *Los ejes X e Y representan respectivamente el desplazamiento (pies) horizontales y verticales respectivamente de los lanzamientos. La velocidad se lanzamiento se fija en el promedio de la muestra para cada tipo de lanzamiento.*

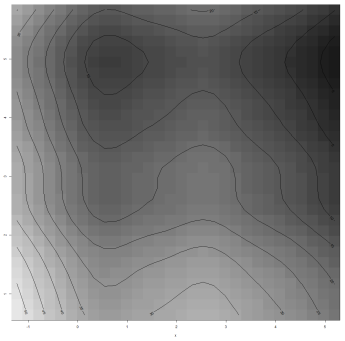
Gráfico 4.8: Zonas de mayor y menor angulo vertical de bateo esperado según el desplazamiento de los lanzamientos en enfrentamientos entre lanzadores y bateadores de distinta lateralidad



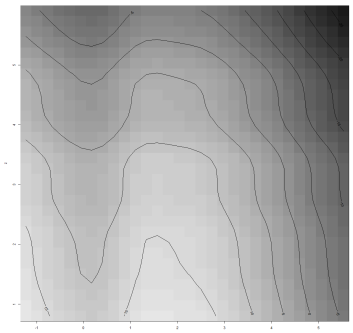
(a) Recta de 4 costuras



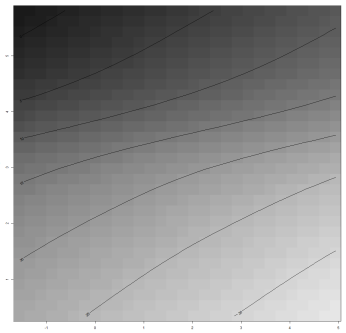
(b) Sinker



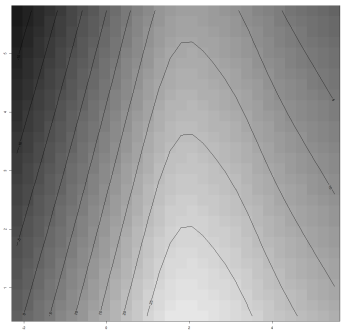
(c) Slider



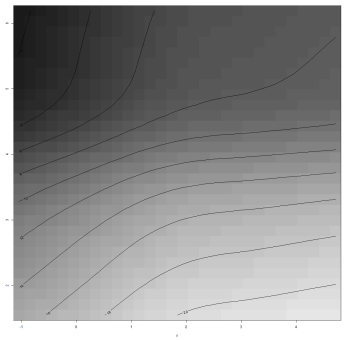
(d) Cambio



(e) Cutter



(f) Sweeper



(g) Curva

Notas: *Los ejes X e Y representan respectivamente el desplazamiento (pies) horizontales y verticales respectivamente de los lanzamientos. La velocidad se lanzamiento se fija en el promedio de la muestra para cada tipo de lanzamiento.*

Capítulo 5

Conclusiones

El estudio ha revelado patrones significativos en la relación entre las características de los lanzamientos y los resultados de los contactos de bateo. Utilizando modelos de regresión spline aditiva, se ha demostrado que los lanzamientos ubicados en las esquinas inferiores de la zona de bateo tienden a generar contactos más débiles, y que el movimiento vertical de los lanzamientos es crucial para inducir batazos rodados. Estos hallazgos están en línea con estudios previos, como el de Nakahara et al. (2023), que encontró que lanzar fuera es más efectivo que lanzar dentro, y el de Yee and Deshpande (2024), que identificó la ubicación del lanzamiento como el factor más importante para predecir el contacto del bateador. Además, el estudio confirma la importancia de la variabilidad el movimiento de los lanzamientos, como se ha discutido en trabajos anteriores de Healey (2019) y Martin (2019).

5.1. Desafíos enfrentados

La investigación enfrentó varios desafíos significativos. En primer lugar, la comprensión y el procesamiento de los datos de "statcast" de MLB requirieron un conocimiento previo detallado y una preparación exhaustiva. La creación de la base de datos para la estimación implicó procesos complejos de extracción y transformación de un gran volumen de datos, utilizando API y el lenguaje R. Este proceso no solo fue intensivo en términos de tiempo, sino que también requirió habilidades técnicas avanzadas para garantizar la precisión y la integridad de los datos.

En términos metodológicos, la selección de la metodología adecuada fue un reto importante. Entre las diversas opciones disponibles, se consideraron y descartaron metodologías como la regresión logística y los modelos de clasificación debido a sus limitaciones para capturar la complejidad de las relaciones no lineales en los datos. Adicionalmente, estas metodologías agruparían los resultados de bateo resultados de bateo particulares, ignorando las diferencias en resultados de bateo entre los individuos de un mismo grupo. La regresión spline aditiva se eligió por su capacidad para modelar relaciones no lineales de manera flexible, mientras que los contornos de los conjuntos de nivel aportaron una visión general de los resultaos de bateo esperados dado los determinantes involucrados.

5.2. Limitaciones del estudio y futuras líneas de investigación

El estudio presenta varias limitaciones que deben ser consideradas. Una de las principales limitaciones es la no inclusión de todos los posibles factores que determinan los resultados de bateo, debido al problema de la dimensionalidad en las estimaciones no paramétricas. Además, los resultados se interpretan como los efectos principales de la localización y el movimiento en los lanzamientos típicos de cada tipo, ya que la velocidad se fijó en el promedio de la muestra para cada lanzamiento y no

se permitió la interacción entre variables. Esto implica que los resultados podrían variar al considerar distintas velocidades o interacciones entre variables explicativas.

Futuras líneas de investigación podrían centrarse en la inclusión de más factores que influyen en los resultados de bateo, así como en el desarrollo de modelos que permitan la interacción entre variables. Además, la aplicación de técnicas de aprendizaje automático y redes neuronales podría proporcionar una comprensión más profunda y detallada de los patrones en los datos.

5.3. Implicaciones prácticas y teóricas

Las implicaciones prácticas de este estudio son significativas para el desarrollo de estrategias de pitcheo más efectivas. Los lanzadores y entrenadores pueden utilizar estos hallazgos para ajustar sus tácticas, enfocándose en las zonas y tipos de movimiento que se ha demostrado que generan contactos más débiles y batazos rodados. Esto podría traducirse en una mejora en el rendimiento defensivo del equipo y una reducción en la cantidad de hits permitidos.

Desde una perspectiva teórica, este estudio contribuye al cuerpo de conocimiento existente sobre la interacción entre lanzadores y bateadores en el béisbol. Al utilizar técnicas avanzadas de análisis de datos, se ha proporcionado una comprensión más detallada y matizada de cómo las características específicas de los lanzamientos influyen en los resultados de los contactos de bateo. Esto abre nuevas vías para futuras investigaciones que podrían explorar otros aspectos del juego, como la influencia del contexto del partido o las características individuales de los jugadores.

5.4. Recomendaciones para la práctica deportiva

Basado en los hallazgos de este estudio, se pueden hacer varias recomendaciones prácticas para los lanzadores y entrenadores de béisbol:

- **Enfocarse en las esquinas inferiores de la zona de bateo**: Los lanzamientos ubicados en estas áreas tienden a generar contactos más débiles, lo que puede reducir la probabilidad de hits exitosos.
- **Aprovechar el movimiento vertical de los lanzamientos**: El movimiento vertical es crucial para inducir batazos rodados, que son más fáciles de manejar para la defensa.
- **Personalizar las estrategias de lanzamiento**: Adaptar las tácticas de lanzamiento según el tipo de enfrentamiento (misma o distinta lateralidad) y el tipo de lanzamiento puede maximizar la efectividad del pitcheo.

En resumen, este estudio ha proporcionado una visión valiosa de la relación entre las características de los lanzamientos y los resultados de los contactos de bateo, a pesar de los desafíos y limitaciones enfrentados. Los hallazgos obtenidos no solo contribuyen al conocimiento científico en el campo del análisis deportivo, sino que también tienen el potencial de transformar las estrategias de entrenamiento y juego en el béisbol.

Bibliografía

Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, and Volkan Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. *Advances in Neural Information Processing Systems*, 29, 2016.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

Brent Bryan, Robert C Nichol, Christopher R Genovese, Jeff Schneider, Christopher J Miller, and Larry Wasserman. Active learning for identifying function threshold boundaries. *Advances in Neural Information Processing Systems*, 18, 2005.

Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Laurent Cavalier. Nonparametric estimation of regression level sets. *Statistics: A Journal of Theoretical and Applied Statistics*, 29(2):131–160, 1997.

David J Cox, Jacob Sosine, and Jesse Dallery. Application of the matching law to pitch selection in professional baseball. *Journal of Applied Behavior Analysis*, 50(2):393–406, 2017.

Hai Dang Dau, Thomas Laloë, and Rémi Servien. Exact asymptotic limit for kernel estimation of regression level sets. *Statistics Probability Letters*, 161:108721, 2020.

Woojin Doo and Heeyoung Kim. Modeling the probability of a batter/pitcher matchup event: A bayesian approach. *Plos One*, 13(10):e0204874, 2018.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*, pages 85–100. Springer, 1977.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

Julian J Faraway. *Linear Models with R.* Chapman and Hall, 2015.

Rob Gray. Markov at the bat: A model of cognitive processing in baseball batters. *Psychological Science*, 13(6):542–547, 2002.

Huong Ha, Sunil Gupta, Santu Rana, and Svetha Venkatesh. High dimensional level set estimation with bayesian neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12095–12103, 2021.

Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993.

Wolfgang Härdle, Axel Werwatz, Marlene Müller, Stefan Sperlich, Wolfgang Härdle, Axel Werwatz, Marlene Müller, and Stefan Sperlich. *Nonparametric Regression*. Springer, 2004.

Glenn Healey. Modeling the probability of a strikeout for a batter/pitcher matchup. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2415–2423, 2015.

Glenn Healey. Matchup models for the probability of a ground ball and a ground ball hit. *Journal of Sports Analytics*, 3(1):21–35, 2017.

Glenn Healey. A bayesian method for computing intrinsic pitch values using kernel density and non-parametric regression estimates. *Journal of Quantitative Analysis in Sports*, 15(1):59–74, 2019.

Glenn Healey and Shiyuan Zhao. Using pitchf/x to model the dependence of strikeout rate on the predictability of pitch sequences. *Journal of Sports Analytics*, 3(2):93–101, 2017.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

Hyunuk Kim and Woo-Sung Jung. Does pitch type-zone uncertainty matter to a pitcher's performance? *New Physics: Sae Mulli*, pages 624–629, 2018.

Thomas Laloë and Rémi Servien. Nonparametric estimation of regression level sets using kernel plug-in estimator. *Journal of the Korean Statistical Society*, 42:301–311, 2013.

Eric P Martin. Predicting major league baseball strikeout rates from differences in velocity and movement among player pitch types. In *MIT Sloan Sports Analytics Conference*, 2019.

Luke McElroy. Computer vision in baseball: The evolution of statcast. *Unknown*.

Joshua Mizels, Brandon Erickson, and Peter Chalmers. Current state of data and analytics research in baseball. *Current Reviews in Musculoskeletal Medicine*, 15(4):283–290, 2022.

Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2012.

Hiroshi Nakahara, Kazuya Takeda, and Keisuke Fujii. Pitching strategy evaluation via stratified analysis using propensity score. *Journal of Quantitative Analysis in Sports*, 19(2):91–102, 2023.

Wolfgang Polonik and Zailong Wang. Estimation of regression contour clusters—an application of the excess mass approach to regression. *Journal of Multivariate Analysis*, 94(2):227–249, 2005.

C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

Clayton Scott and Mark Davenport. Regression level set estimation via cost-sensitive classification. *IEEE Transactions on Signal Processing*, 55(6):2752–2757, 2007.

Shubhanshu Shekhar and Tara Javidi. Multiscale gaussian process level set estimation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3283–3291. PMLR, 2019.

Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

Rebecca M Willett and Robert D Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.

Ryan Yee and Sameer K Deshpande. Evaluating plate discipline in major league baseball with bayesian additive regression trees. *Journal of Quantitative Analysis in Sports*, 20(1):5–20, 2024.

Glosario

Barrel: La parte más gruesa del bate, el punto ideal para conectar la pelota y obtener la máxima potencia. Statcast clasifica con este término los contactos con combinación optima de velocidad y ángulos de salida.

Bate: Instrumento de madera o metal que utilizan los bateadores para golpear la pelota.

Breaking (ball): Término general para cualquier lanzamiento que se curva o se mueve de manera impredecible:, dificultando el bateo.

Cambio: Un lanzamiento que parece una recta rápida pero se rompe hacia abajo en el último momento.

Cuadrangular (Home run): Un hit que permite al bateador recorrer todas las bases y anotar una carrera.

Curva: Un lanzamiento que se curva hacia abajo y lateralmente.

Cutter: Lanzamiento rápido que se mueve lateralmente, similar a un slider pero con menos curvatura.

Doble: Un hit que permite al bateador llegar a segunda base.

Eephus: Un lanzamiento extremadamente lento utilizado para sorprender al bateador. Poco utilizado en la practica.

Elevado: Un batazo elevado que viaja por el aire en un ángulo pronunciado en relación con el terreno.

Extrabase: Un hit que permite al bateador llegar más alla de la primera base.

Fair Ball: Una pelota bateada que, al ser golpeada por el bate, viaja dentro de las líneas de foul, que se extienden desde el home plate hasta los postes de foul atravezando la primera y tercera base.

Fastball: Un lanzamiento rápido y recto, la base de la mayoría de los repertorios de pitcheo.

Hawk-Eye: Sistema de seguimiento de la pelota que utiliza cámaras de alta velocidad para rastrear su trayectoria.

Hit: Un batazo que pone en juego la pelota y permite al bateador alcanzar una base de manera segura.

Línea: Un batazo solido y rasante que viaja a gran velocidad.

MLB (Major League Baseball): Las Grandes Ligas de Béisbol, la máxima categoría del béisbol profesional en Estados Unidos y Canadá.

Nudillo (Knuckleball): Un lanzamiento que se mueve de manera errática debido a la rotación irregular de la pelota.

Offspeed: Término general para cualquier lanzamiento que sea más lento que la recta rápida.

Out: Cuando un corredor es eliminado o un bateador es puesto fuera.

Ponche (Strike out): Se refiere al momento en el que un bateador es eliminado al acumular tres strikes.

Pop up: Un elevado corto y fácil de atrapar para un infielder.

Recta: Otro término utilizado para fastball, un lanzamiento rápido y recto.

Rodado (Ground ball): Un batazo que golpea el suelo y se desplaza a lo largo del terreno.

Screwball: Un lanzamiento que se mueve lateralmente en dirección opuesta a un cutter, aunque es menos común.

Sencillo: Un hit que permite al bateador llegar a primera base.

Sinker: Un lanzamiento rápido que se hunde hacia abajo al final de su trayectoria.

Slider: Un lanzamiento que se mueve lateralmente, similar a un cutter pero con más curva.

Slurve: Un híbrido entre una curva y un slider, con una trayectoria más suave y menos definida.

Split-Finger: Un lanzamiento rápido con dos dedos separados en la pelota, lo que provoca una caída brusca al final de su trayectoria.

Statcast: Un sistema de seguimiento de datos que recopila información detallada sobre cada jugada en un juego de béisbol.

Strike: Es un lanzamiento que el árbitro considera que el bateador "debería" haber conectado (dentro de la zona de strike o Swing fallado). Si un bateador acumula tres strikes, es eliminado (ponchado). .

Sweeper: Un lanzamiento que se mueve lateralmente con un arco más amplio que un slider, describiendo una trayectoria similar a la de una escoba barriendo el suelo.

sweet spot: Es una pequeña área en la parte más gruesa del bate (barrel). En este punto las vibraciones se minimizan, lo que permite al bateador sentir el impacto de manera más sólida y enviar la pelota más lejos.

Swing: EL movimiento de balancear el bate para golpear la pelota.

Tenedor (Forkball): Similar al split-finger, pero con los dedos más juntos.

Triple: Un hit que permite al bateador llegar a tercera base.

wOBA (Weighted On-base Average): Estadística avanzada que busca medir el valor ofensivo de un jugador asignando un valor a cada evento ofensivo (sencillos, dobles, triples, jonrones, bases por bolas, etc.) en función de su impacto real en la probabilidad de anotar una carrera.

Anexos

Anexo 1: Clasificación de lanzamientos en Statcast y variaciones de nombre en español

Grupo	Subgrupo	Lanzamiento	Nombre en español
Fastball		Fastball (4-seam)	Recta de 4 costuras
		Sinker (2-Seam)	Recta de 2 costuras
		Cutter	Recta cortada
Offspeed		Changeup	Cambio
		Split-finger	Recta de dedos separados
		Forkball	Tenedor
		Screwball	-
Breaking	Curveball	Curveball	Curva
		Knuckle Curve	Curva de nudillos
		Slow Curve	Curva lenta
	Slider	Slider	-
		Sweeper	-
		Slurve	-
	Knuckleball	Knuckleball	nudillos
Otros lanzamientos		Eephus	-
		Intentional Ball	Bola intencional
		Pitchout	-
		Otros	-

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*

Anexo 2: Pruebas de independencia de Pearson entre la "familia del tipo de lanzamiento" y variables categóricas seleccionadas

Variable	**Chi-cuadrado**	**Grados de libertad**	**P-valor**
Barrel	108,17	2	2,20E-16
Contacto fuerte	1.156,6	2	2,20E-16
Ángulo de salida	426,62	6	2,20E-16

Fuente: *Elaboración propia del autor a partir de los datos de Statcast*
Notass: *La variable "tipo de lanzamiento" contiene las categorías: recta, offspeed y rompiente. La variable "ángulo de salida" agrupa; rodado, línea, elevado y popup.*

Anexo 3: Pruebas de Kruskal-Walllis de igualdad en distribución de variables continúas seleccionadas entre categorías de definidas por el ángulo de salida

Variable	Chi-cuadrado	Grados de libertad	P-valor
Velocidad	65,116	3	4,737e-14
Localización horizontal	1584,5	3	4,737e-14
Localización vertical	4942,1	3	2,2e-16
Desplazamiento horizontal	182,52	3	2,2e-16
Desplazamiento vertical	2401,1	3	2,20e-16

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *La clasificación de "ángulo de salida" incluye las categorías; elevado, línea, rodado, y popup.*

Anexo 4: Pruebas de Kruskal-Walllis de igualdad en distribución de variables continúas seleccionadas entre categorias de tipo de enfrentamiento

Variable	Chi-cuadrado	Grados de libertad	P-valor
Velocidad	1.625	3	2,20e-16
Localización horizontal	211,84	3	2,20e-16
Localización vertical	65,83	3	3,34e-14
Desplazamiento horizontal	82,194	3	2,20e-16
Desplazamiento vertical	228,87	3	2,20e-16
Velocidad de salida	165,83	3	2,20e-16
Angulo de salida	235,18	3	2,20e-16

Fuente: *Elaboración propia del autor a partir de los datos de Statcast.*
Notas: *La variable "tipo de enfrentamiento" contiene las categorias; DD, DI, II y ID, según la combinación de brazo de lanzar del lanzador y lado de bateo del bateador (D: derecho, I: izquierdo).*