Exploring Shot Difficulty on the PGA Tour Using Generalized Additive Modeling and Hierarchical Effects

Benedict Brady

Harvard Sports Analysis Collective

April 30, 2019

History

- Early golf metrics involved how players scored relative to par, putts per hole, or greens in regulation.

History

- Early golf metrics involved how players scored relative to par, putts per hole, or greens in regulation.
- In the early 2000s, the PGA Tour started recording X, Y, Z data for essentially all of the shots taken on the Tour since 2000 ($\sim$ 2 million per year).

History

- Early golf metrics involved how players scored relative to par, putts per hole, or greens in regulation.
- In the early 2000s, the PGA Tour started recording X, Y, Z data for essentially all of the shots taken on the Tour since 2000 ($\sim$ 2 million per year).
- Around 2010, researchers started getting interested in using shot difficulty to infer player strength.

History

- Early golf metrics involved how players scored relative to par, putts per hole, or greens in regulation.
- In the early 2000s, the PGA Tour started recording X, Y, Z data for essentially all of the shots taken on the Tour since 2000 ($\sim$ 2 million per year).
- Around 2010, researchers started getting interested in using shot difficulty to infer player strength.
- The most famous example of this is Broadie 2012 who developed the Strokes Gained metric.

Motivation

- This problem incorporates a large number of features and a large amount of observations.

Motivation

- This problem incorporates a large number of features and a large amount of observations.
- There is almost no repeatability because players do not take shots with similar features very often, especially when distances are incorporated.

Motivation

- This problem incorporates a large number of features and a large amount of observations.
- There is almost no repeatability because players do not take shots with similar features very often, especially when distances are incorporated.
- Metrics that consider stroke level data use naive methods to evaluate shot probability (primarily one or two variable regressions without good solutions for non-linearity).

Motivation

- This problem incorporates a large number of features and a large amount of observations.
- There is almost no repeatability because players do not take shots with similar features very often, especially when distances are incorporated.
- Metrics that consider stroke level data use naive methods to evaluate shot probability (primarily one or two variable regressions without good solutions for non-linearity).
- Sparse features are a good candidate for Hierarchical Models and non-linear data is often best fit using Generalized Additive Modeling.

Data

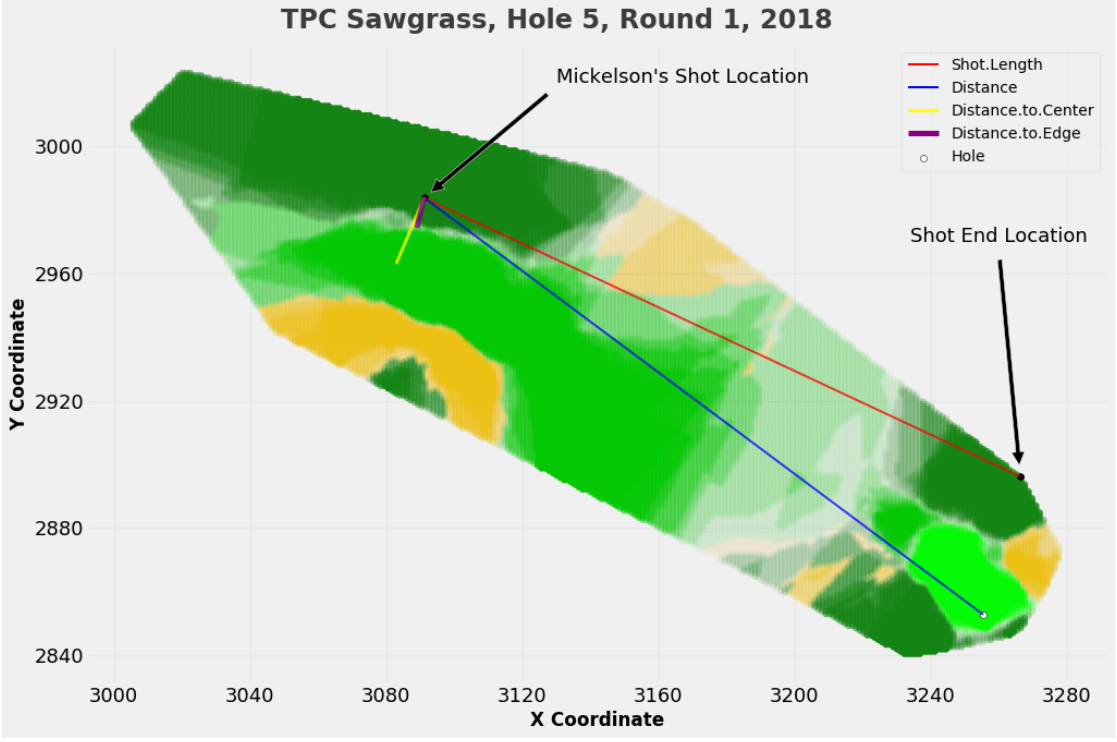
- Main columns I used included Player, Course, Hole, Round, Distance, Location, and Time.

Data

- Main columns I used included Player, Course, Hole, Round, Distance, Location, and Time.
- Engineered features such as hole location, tee location, and effective green.

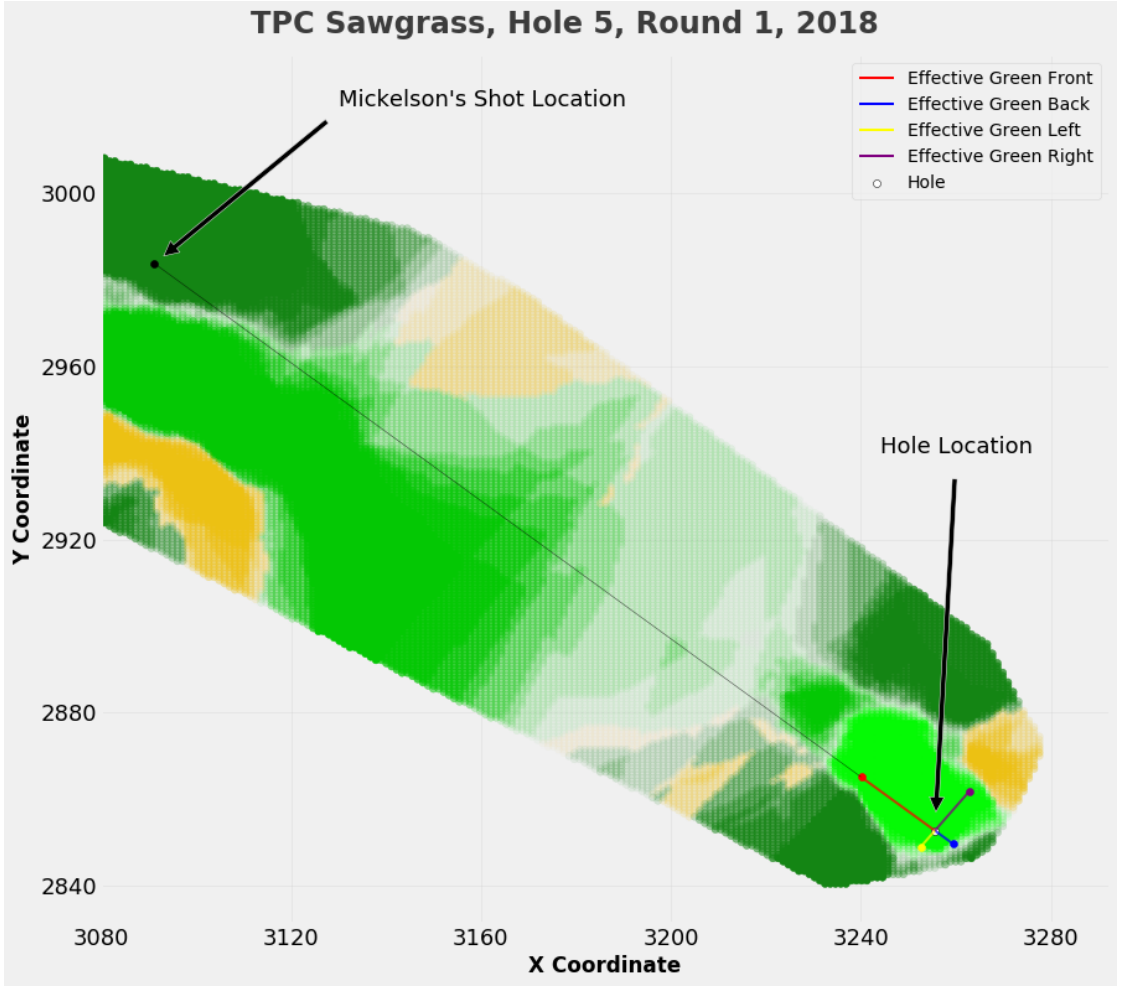
Data

- Main columns I used included Player, Course, Hole, Round, Distance, Location, and Time.
- Engineered features such as hole location, tee location, and effective green.



Effective Green

Effective green is the measure of how much green the golfer has to work with on a given shot.



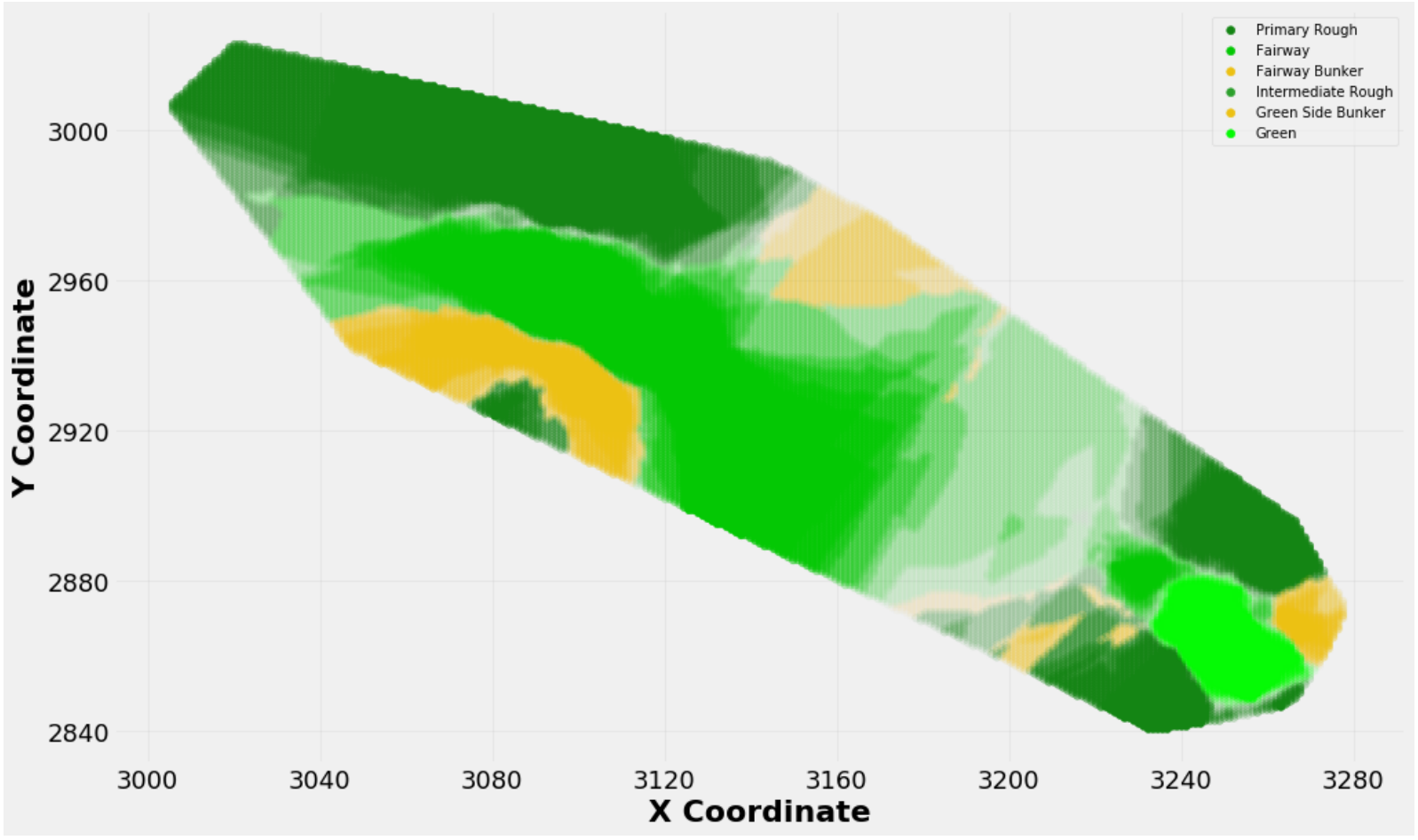
Course Visualizations

- The data provides features for every shot recorded, but this makes out of sample prediction difficult.

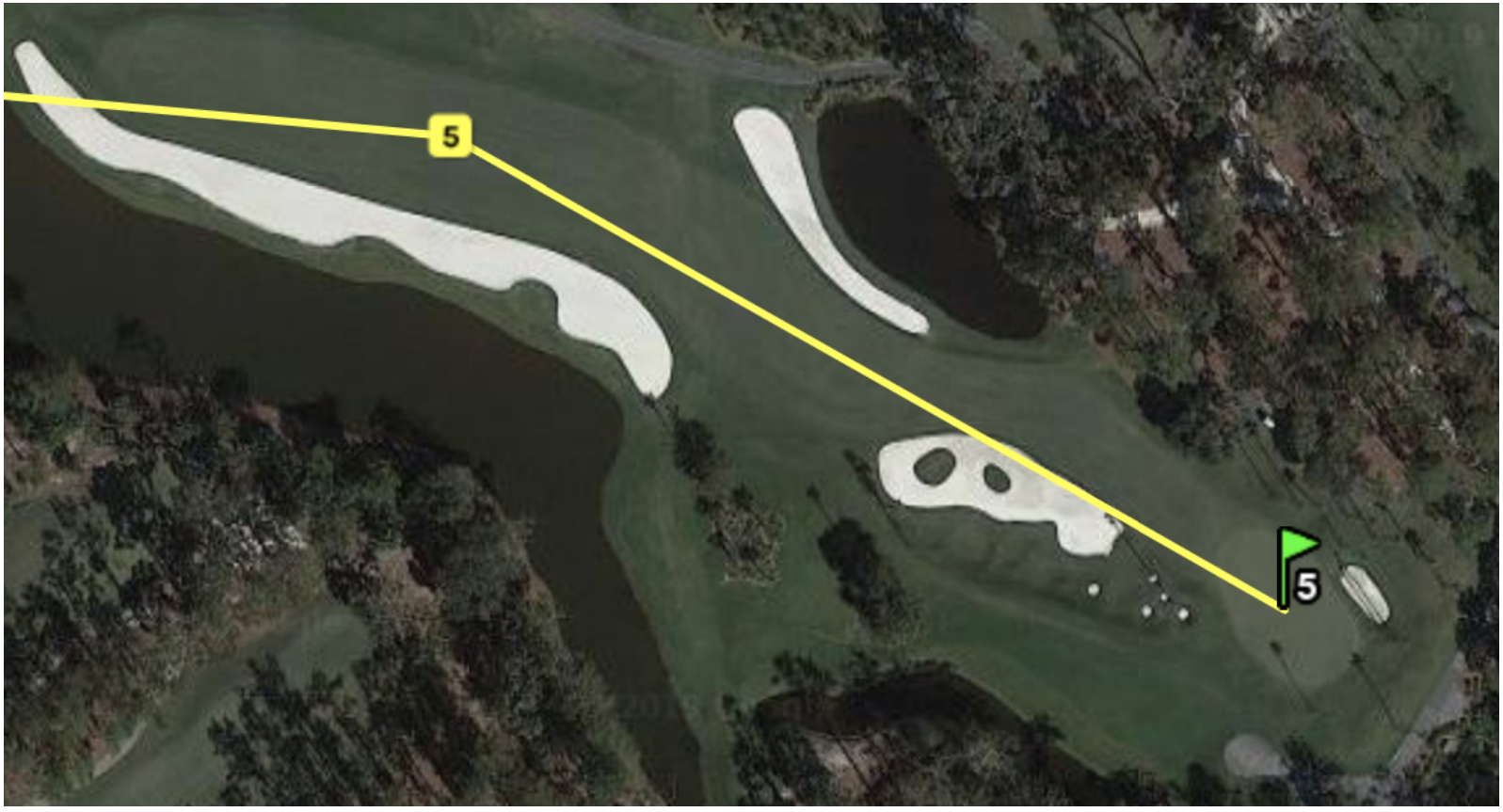
Course Visualizations

- The data provides features for every shot recorded, but this makes out of sample prediction difficult.
- The difficult feature to impute is location. To do this I fit a K Nearest Neighbors model and assigned opacity based on confidence.

Course Viz Example



Course Viz Example

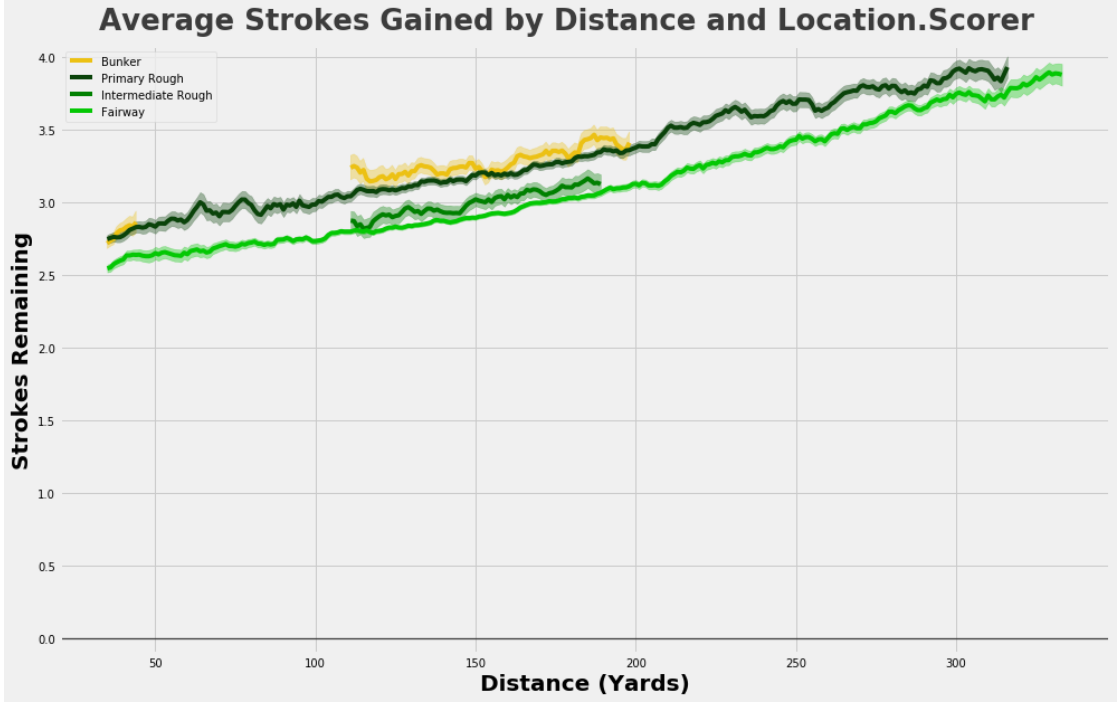


Exploratory Data Analysis

- Most of the stroke level prediction accuracy comes from properly understand the relationship between distance, location type, and shot difficulty.

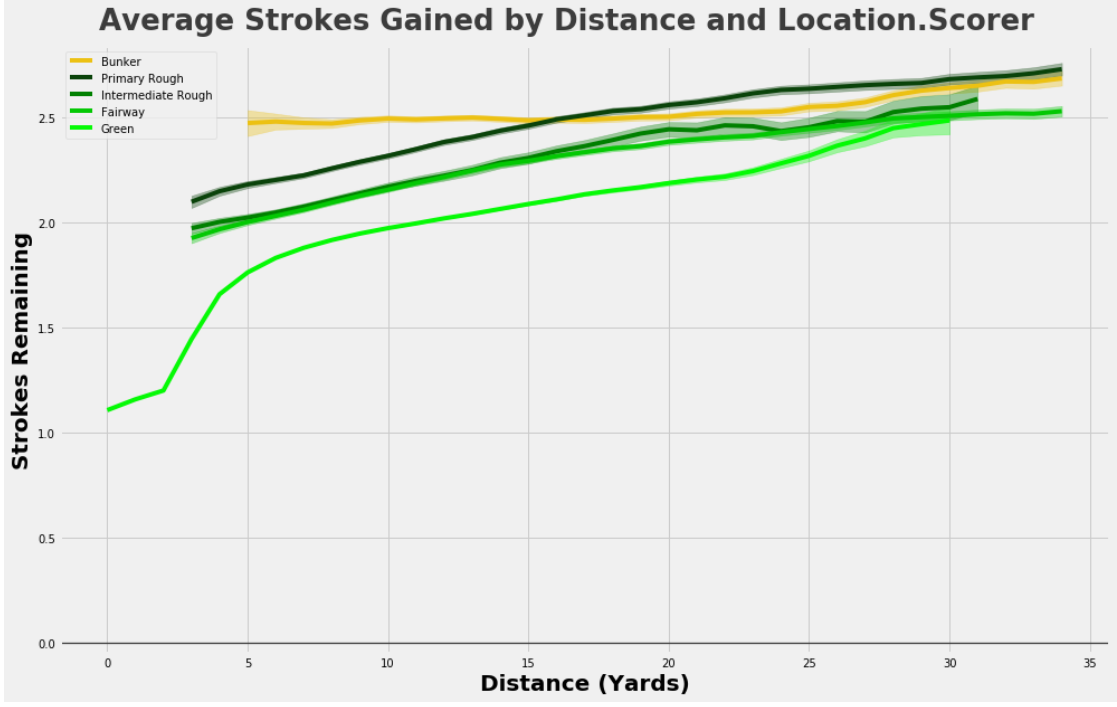
Exploratory Data Analysis

- Most of the stroke level prediction accuracy comes from properly understand the relationship between distance, location type, and shot difficulty.



Exploratory Data Analysis

- Most of the stroke level prediction accuracy comes from properly understand the relationship between distance, location type, and shot difficulty.



Models

- I fit an iterative series of models starting with linear models and introducing both random effects and smooth terms.

Models

- I fit an iterative series of models starting with linear models and introducing both random effects and smooth terms.
- The final models I used were tuned over a large number of parameters in the Generalized Additive Mixed Model structure.

Models

- I fit an iterative series of models starting with linear models and introducing both random effects and smooth terms.
- The final models I used were tuned over a large number of parameters in the Generalized Additive Mixed Model structure.
- These models were high dimensional and needed to be fit using amazon instances and speed-based model optimizations.

Models

- I fit an iterative series of models starting with linear models and introducing both random effects and smooth terms.
- The final models I used were tuned over a large number of parameters in the Generalized Additive Mixed Model structure.
- These models were high dimensional and needed to be fit using amazon instances and speed-based model optimizations.

$$\begin{aligned} &\texttt{S.R} \sim \texttt{s(Dist)} + \texttt{Loc.S} + \texttt{s(Dist, by = Loc.S)} \\ &+ (1|\texttt{Course/Hole/Round}) + (1|\texttt{Player/Loc.S}) + \texttt{s(Time, k = k)} \\ &\quad + \texttt{s(Dist, by = Loc.S)} + \texttt{ti(Time, Dist)} \quad (\texttt{GAM.Time.Dist}) \end{aligned}$$

Model Interpretation

- Using these models I looked at the factors the affected shot difficulty by decomposing the models into basis functions.

Model Interpretation

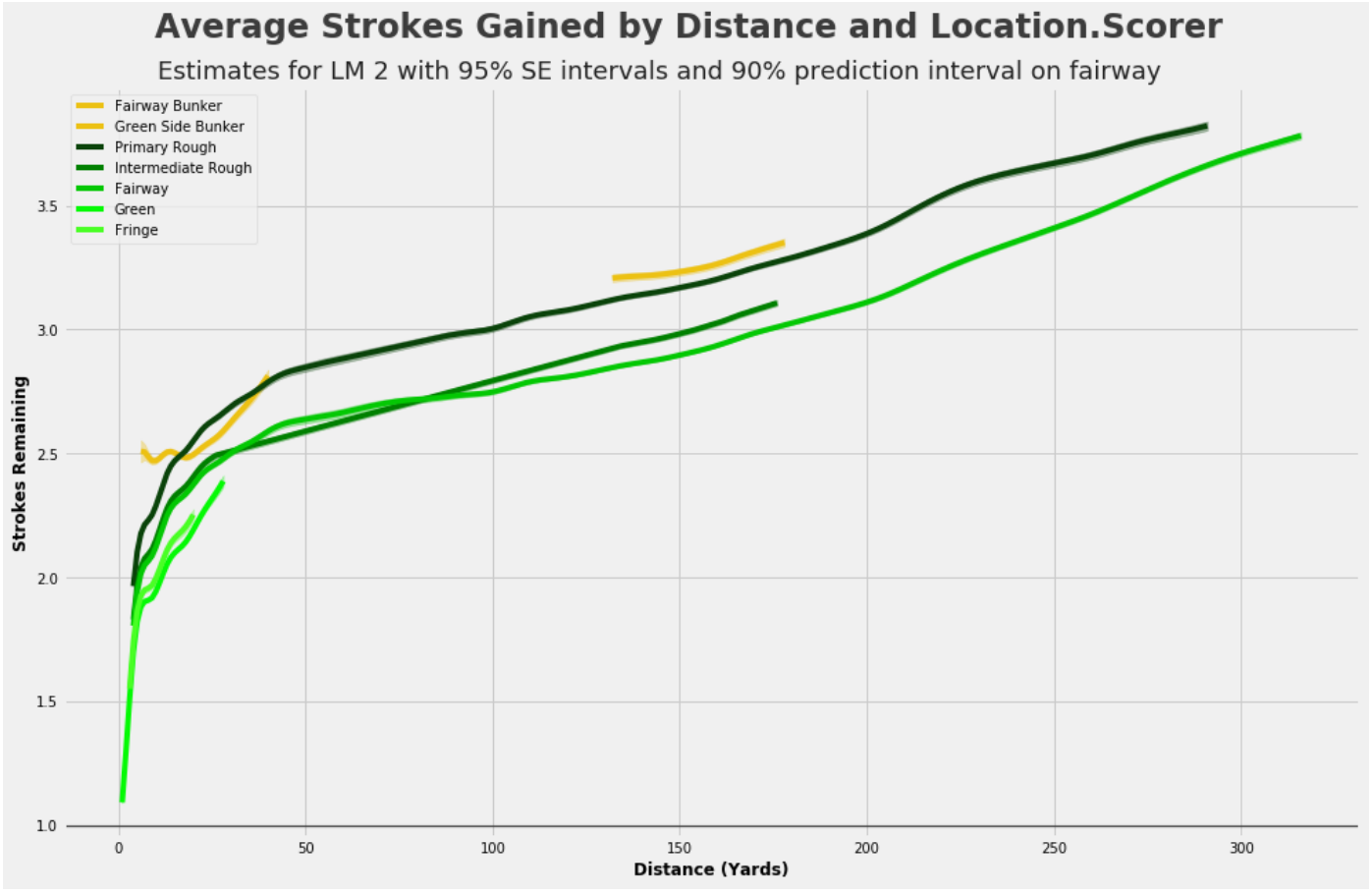
- Using these models I looked at the factors the affected shot difficulty by decomposing the models into basis functions.
- Can also look at player, course and hole rankings for 2018.

Model Interpretation

- Using these models I looked at the factors the affected shot difficulty by decomposing the models into basis functions.
- Can also look at player, course and hole rankings for 2018.
- Finally, I can look at maps of shot difficulty using the course visualization techniques outlined above.

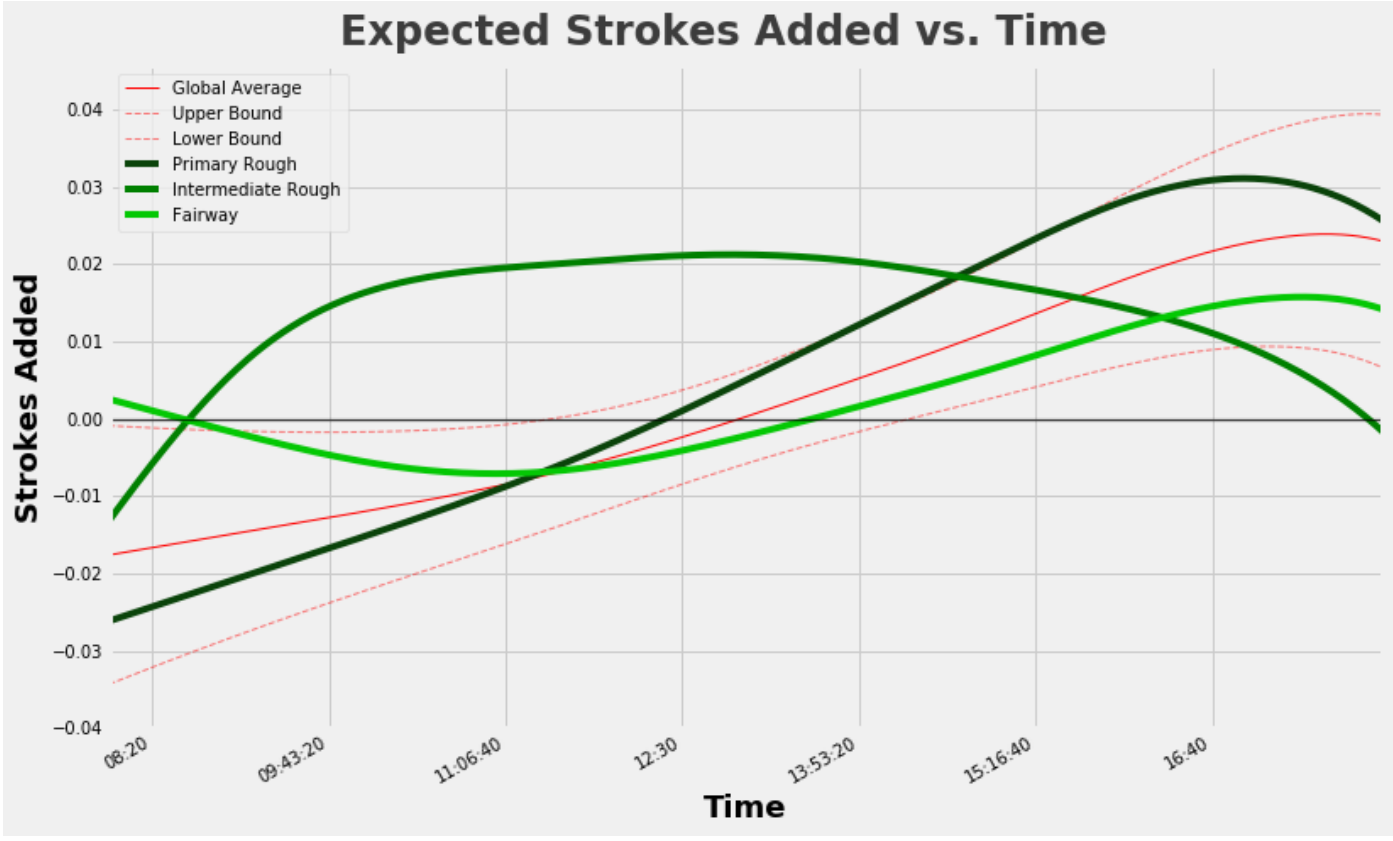
Location and Distance

The first thing to check is the strokes added by distance and location using the smoothing splines.



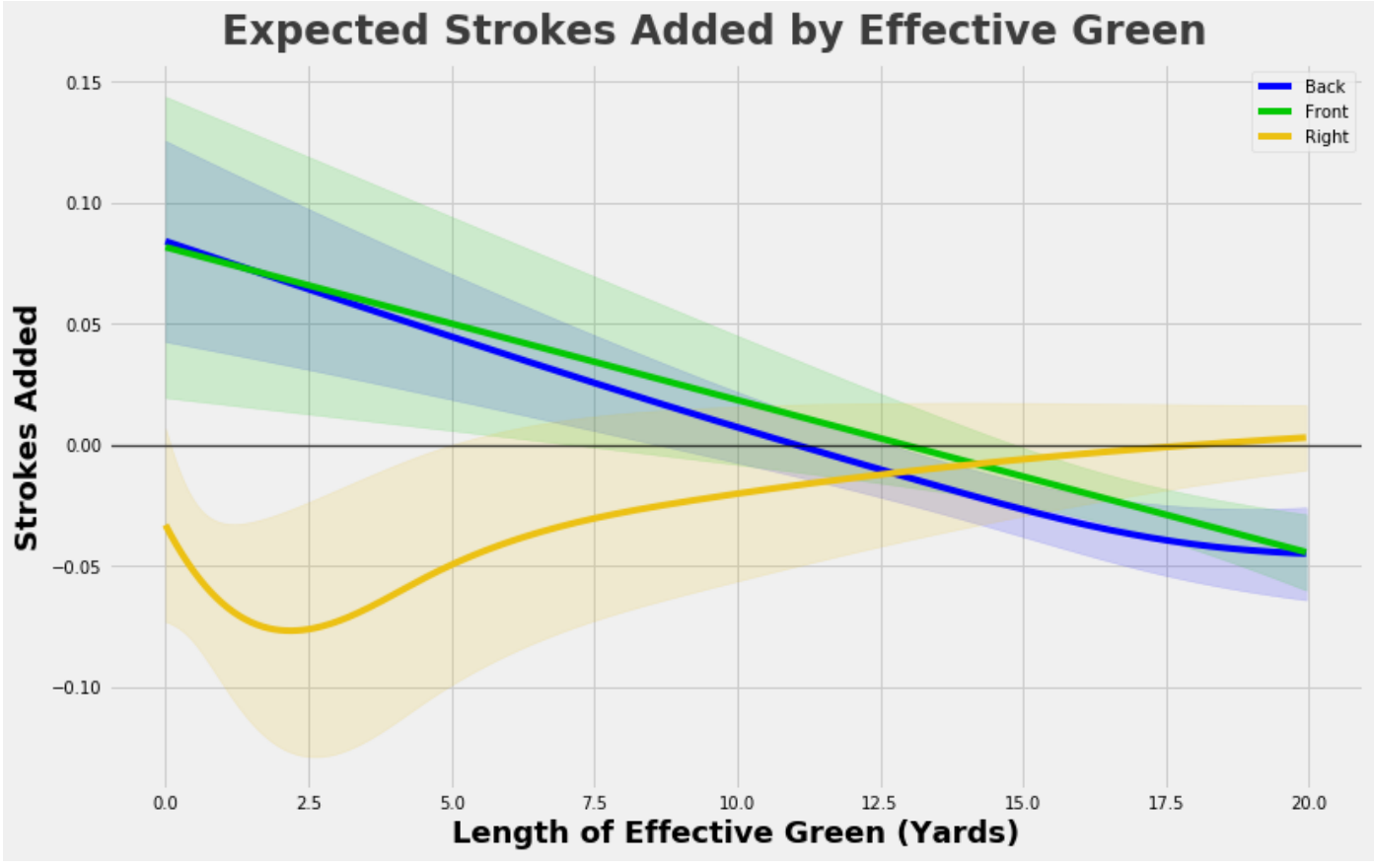
Time of Day

This shows convincing evidence that time of day matters, and that it is more pronounced on shots where it is harder to get spin on the ball.



Effective Green

More effective green is better in the front and the back, with some weird behavior on the right side.



Player Rankings

Can rank the strokes added for each player by location type in 2018.

Green		Fairway		Primary Rough		Bunker	
Player	S.A.	Player	S.A.	Player	S.A.	Player	S.A.
S. Burns	-0.036	T. Woods	-0.10	J. Thomas	-0.064	K. Na	-0.084
G. Chalmers	-0.034	W. Bryan	-0.078	W. Simpson	-0.063	R. Fowler	-0.075
J. Day	-0.030	W. Simpson	-0.076	J. Lovemark	-0.061	J. Day	-0.069
D. Summerhays	-0.029	K. Na	-0.075	P. Reed	-0.061	S. Power	-0.066
W. Simpson	-0.028	J. Day	-0.066	J. Dahmen	-0.060	O. Schniederjans	-0.066
K. Kisner	-0.027	J. Rose	-0.064	T. Hatton	-0.058	A. Baddeley	-0.062
B. Hossler	-0.027	T. Fleetwood	-0.061	B. Haas	-0.052	W. Kim	-0.061
P. Malnati	-0.025	J. Thomas	-0.061	P. Mickelson	-0.050	J. Spieth	-0.060
A. Noren	-0.024	D. Johnson	-0.060	D. Bozzelli	-0.050	D. Lee	-0.060
P. Rodgers	-0.022	P. Mickelson	-0.056	M. Kuchar	-0.048	W. Simpson	-0.057

Course Rankings

Can rank the strokes added for each course on the PGA Tour in 2018.

Course	Strokes Added
Pebble Beach GL	.039
Plantation Course at Kapalua	.031
Silverado Resort and Spa North	.028
Quail Hollow Club	.023
Torrey Pines GC (South)	.023
TPC San Antonio - AT&T Oaks	.022
TPC Summerlin	.022
Riviera CC	.018
Muirfield Village GC	.014
TPC Sawgrass	.012

Hole Rankings

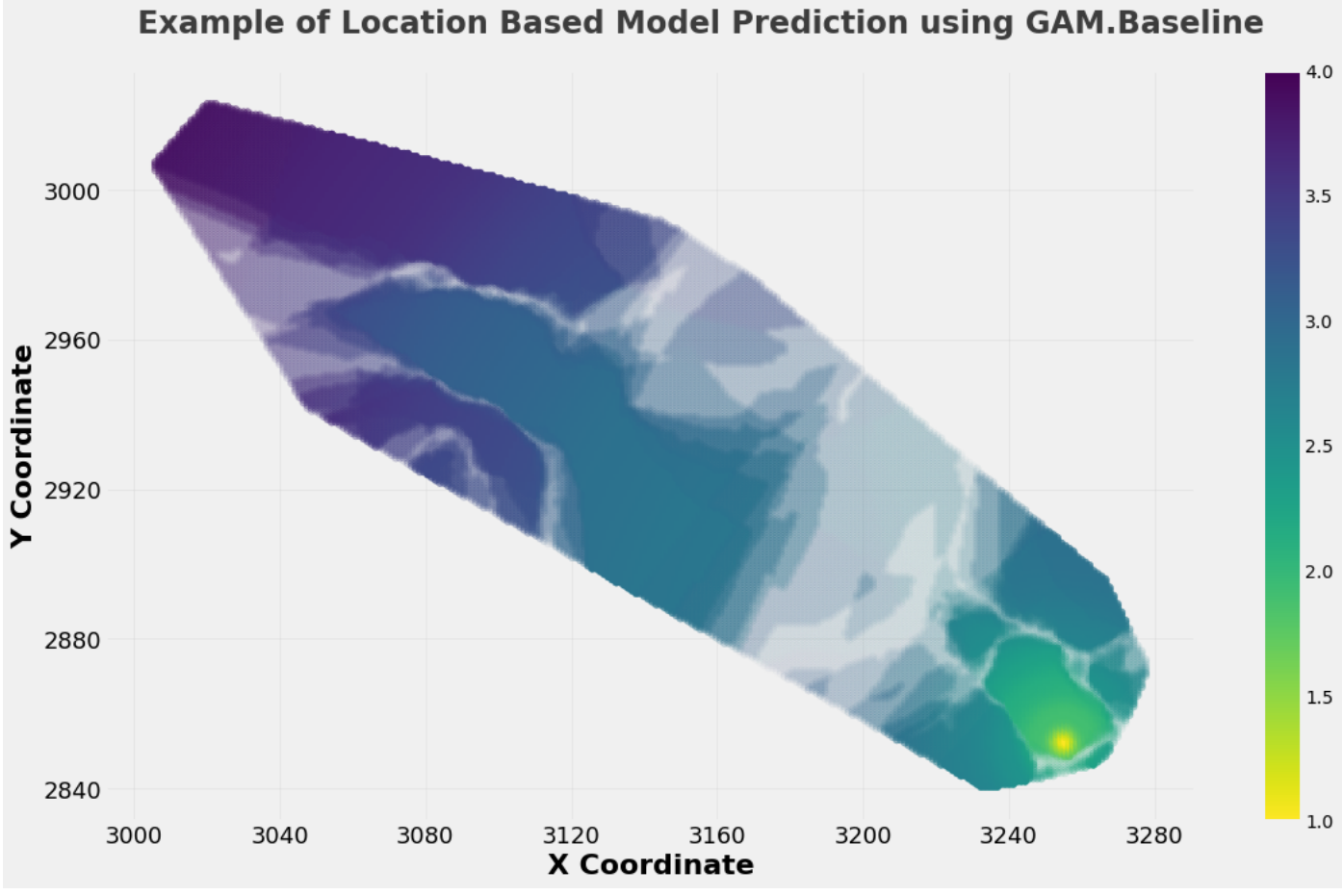
Can rank the strokes added for each course on the PGA Tour in 2018.

Course	Hole	Strokes Added
Riviera CC	10	0.16
Pebble Beach GL	18	0.14
Pebble Beach GL	14	0.13
TPC San Antonio - AT&T Oaks	1	0.12
Torrey Pines GC (South)	13	0.11
Silverado Resort and Spa North	8	0.10
TPC River Highlands	15	0.10
Pebble Beach GL	11	0.09
TPC Summerlin	11	0.08
Silverado Resort and Spa North	3	0.08

Pebble Beach Hole 18

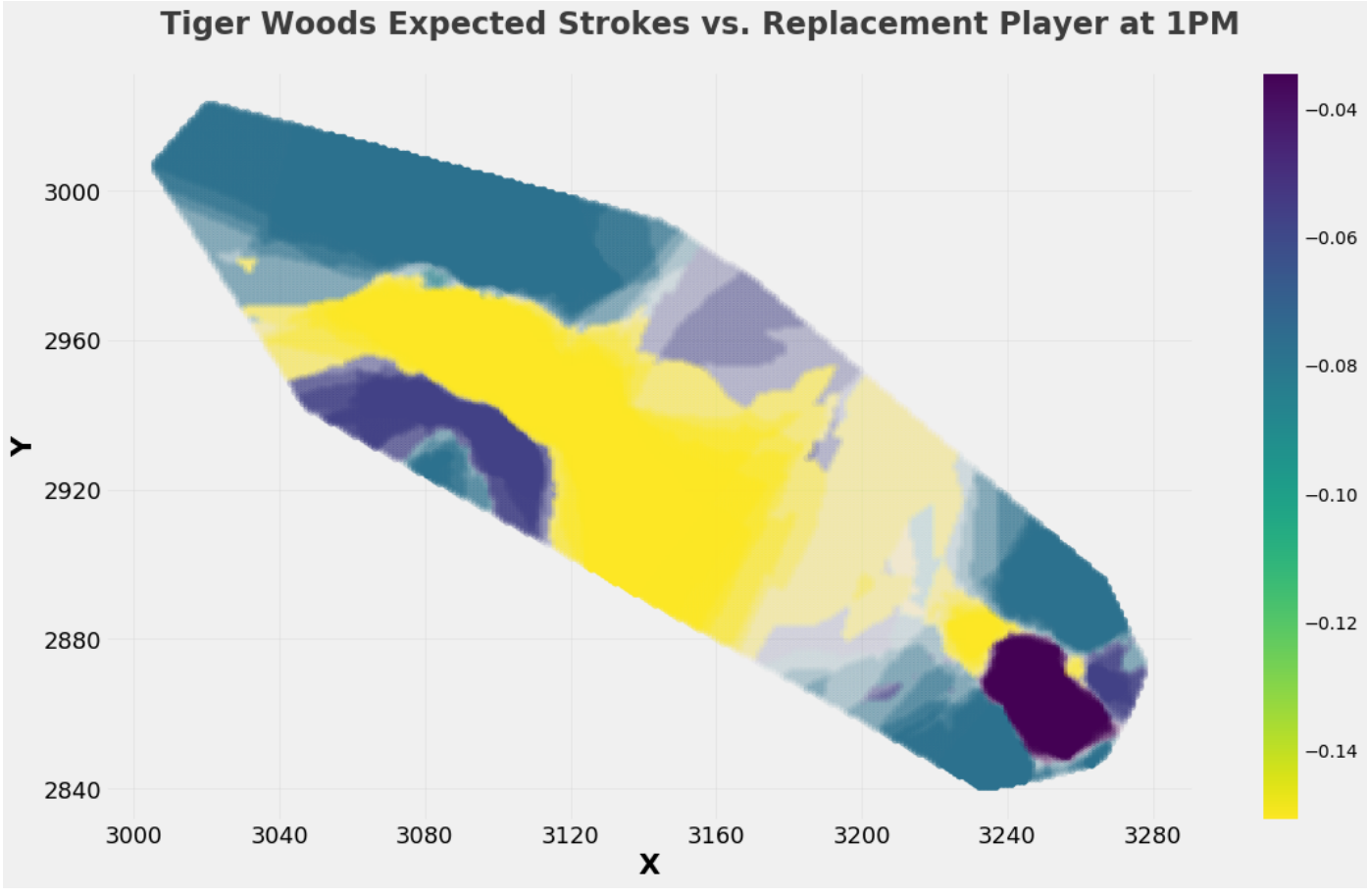
GAM Baseline Predictions

Using the out of sample imputation I mapped the difficulty of the course using the baseline GAM model.



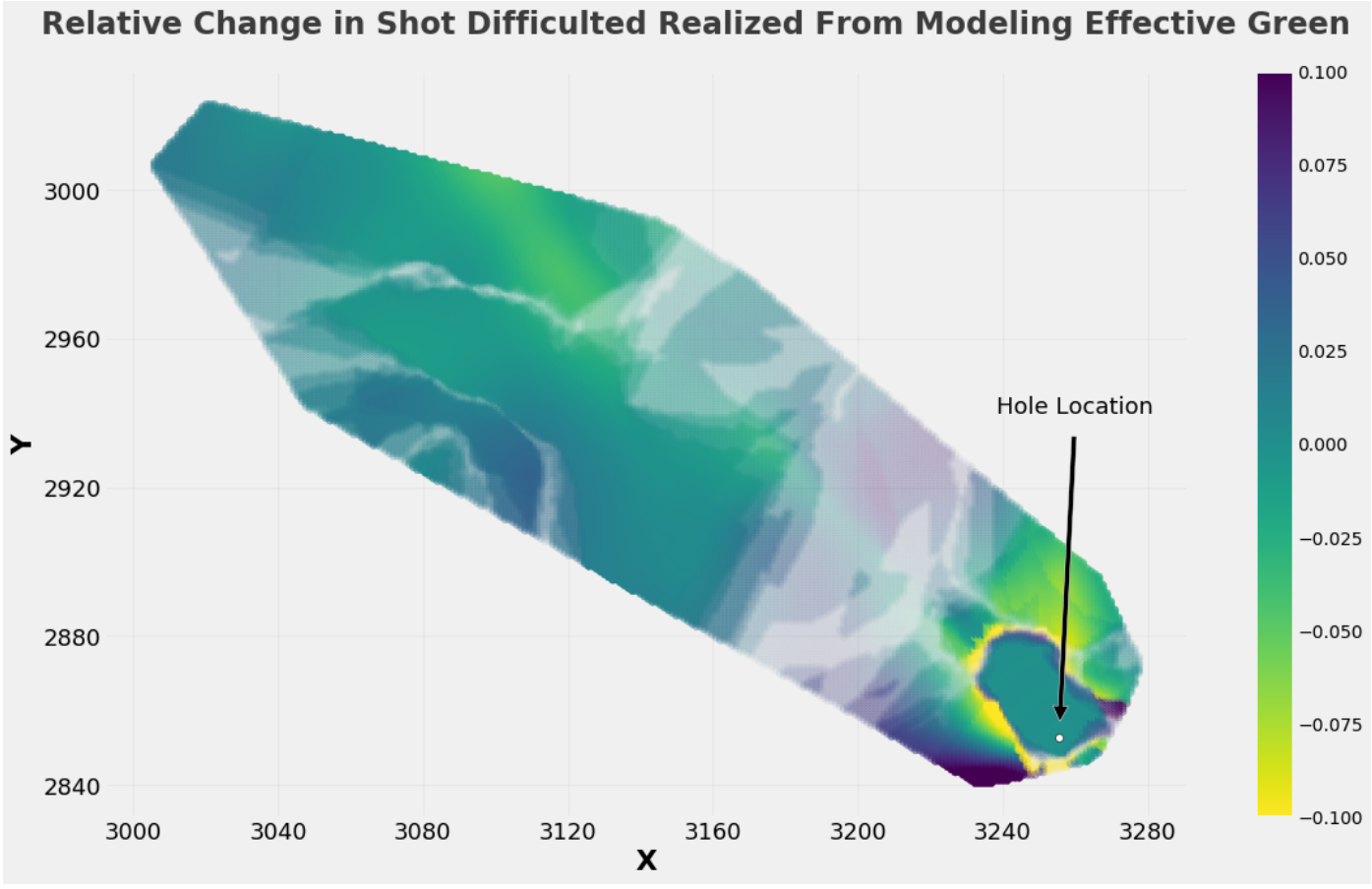
Tiger Woods Strokes Above Baseline

This model makes it simple to compare Tiger Woods to a replacement player on any given hole.



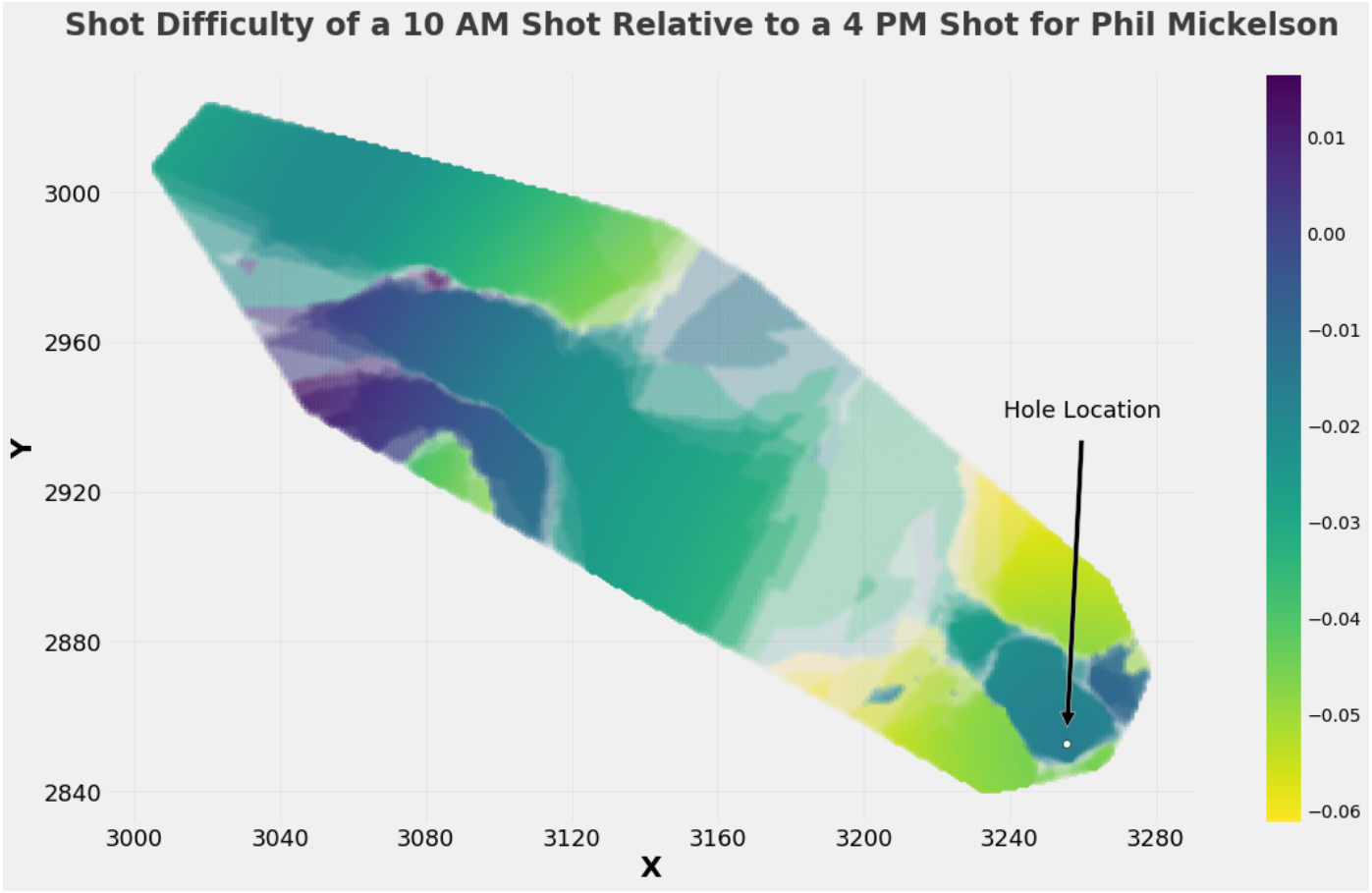
Stroke Difference With Effective Green

This graphic shows the difference in shot difficulty when considering effective green.



Difficulty at 10AM vs. 4PM

Can compare the increase in shot difficulty for Phil Mickelson between 4PM and 10AM.



Conclusion

- This model lays the foundation for a more rigorous way to investigate the impact of specific features on shot difficulty.

Conclusion

- This model lays the foundation for a more rigorous way to investigate the impact of specific features on shot difficulty.
- Can stratify player strength by more granular shot types, or consider more types of location features.

Conclusion

- This model lays the foundation for a more rigorous way to investigate the impact of specific features on shot difficulty.
- Can stratify player strength by more granular shot types, or consider more types of location features.
- Thanks for listening.