How To Break Down a Set Defence

Twenty3 Sport

David Perdomo Meza Daniel Girela Mark Thompson James Goldring

October 7, 2019

Abstract

In football, attacking a team which has transitioned into a deep defensive setup is a common game situation which coaches and players want to understand and plan for. However, despite its practical importance, there hasn't been much research into this situation through the use of event data. In this document, we present an attempt to do precisely that. Using event data supplied by StatsBomb, we present a scalable algorithm to identify events in the game in which a team is attacking against a set defence, explore the empirical outcome of said game situations in the data, and present several approaches to isolate the reward and risk effects of individual factors related to the attacking teams actions when facing an opponent in a set defence.

1 Introduction

Football – with its eleven players on each team on the pitch – is an incredibly dynamic sport with endless variations. And yet, the nature of this kind of goal-scoring sport (where points are gained by getting the ball past a goalkeeper into a goal) makes it a sensible, and frequently seen, strategy for defending teams to keep as many players between themselves and the goal that they're defending as possible, to prevent the attacking team's route to goal.

Because of this, attacking teams will often find themselves faced with an opponent who are set in a compact defensive shape, usually just on the edge of the eighteen-yard box, blocking the route to goal. Trying, and struggling, to break down an opponent who is set in their defensive shape is a situation which teams will find themselves in frequently, and in every match that they play. Often, they will fall into what we refer to in the paper as *probing* — retaining possession of the ball but passing it amongst themselves in a not-altogether-fast manner which, to an outside observer, can look a little aimless.

There are obvious reasons to want to improve on the situation described above, and the low-scoring nature of the sport means that finding a way to break down these set defensive structures with even slightly more efficacy would be a great boost to football teams.

With the importance for practitioners and coaches to understand and plan for these crucial game situations, it is only natural to look towards data analytics as a valuable tool in providing insight. Increasingly so amongst the world's most popular sports, the raw volume of data collection, coupled with the sophistication of techniques to interrogate it for answers and insight, has put data analytics at the forefront of many workflows within professional sports clubs.

However, when it comes to providing insight into these crucial game situations of attacking against a set defence through the use of data, the conceptual definitions above do not have a natural equivalent in the data, particularly event data. As opposed to more segmented, stop-start sports like American football or baseball, game situations in football are rarely neatly categorised and labelled within data feeds. As is recurrent in football analytics, the study of these situations using data stumbles upon a challenge from conception, with the lack of established and formal definitions of what we want to analyse.

This area of study is not ground that is well-trodden. There is little work on defining phases of play beyond set-pieces and counter-attacks, the principal of which in the public sphere is Opta Pro's work [1], although their methodology itself has not been made public.

In this document, we present original research which attempts to provide some data-driven insight into how to break down a set defence. We present an end-to-end study; from proxying the game situation in the event data, through to actionable insight into how to successfully attack a set defence. The document is organised as follows:

- In Section 2 we concern ourselves with the proxy for *attacking a set defence*, describing how we designed the criteria to determine when we consider that an attacking event occurred against a defence that was set.
- Using this definition, in Section 3 we carry out an empirical examination of the results of teams attacking a set defence, beginning to identify the different locations from which an attack against a set defence is successful when compared to other types of attacks, and begin to beckon to the practitioner with some practical and digestible results concerning the success of these attacks.
- In Section 4 we take this exercise even further, fully breaking down the outcome of individual factors of attacks against a set defence. We present two separate approaches to deal with different segments of possessions which are against a set defence; and evaluate the respective risk and reward of each as per the legend below.

	Reward	**Risk**
Events which we consider to occur against a set defence	Section 4.1	Section 4.3.1
Events in which a team faced a "set" defence at some point, but not currently	Section 4.2	Section 4.3.2

- Finally, in Section 5 we conclude and offer some remarks for future work.

2 Defining Attacking Events Against a Set Defence

The first challenge we must face is the proxying of attacking a set defence in the available data. For this project we used event data supplied by Statsbomb for the 2018/19 season from the Premier League, Bundesliga, La Liga, Serie A, and Ligue 1. "Event data" provides one observation for each on-field event (as opposed to "tracking data" which gives the location of all 22 players, as well as the ball, for each moment of the match).

StatsBomb's data provision, as well as the basic information about on-field events, included a few other features which we utilised in our analysis: each event was assigned to a **possession**, making it easy to analyse

sequences of possession rather than just each event as its own isolated action. Events were therefore given information of which team was in possession, and the individual possession changed if and when play was restarted or their opponent established control of the ball. Events were also assigned a **play pattern**, noting whether the possession they were part of came in regular play, from a set-piece (of various kinds), or in a counter attack. We found this designation of "Regular Play" and "From Counter" particularly useful during the analysis.

The type of available data, on the ball events, ensures that proxying a set defence is particularly challenging because there is a natural bias towards *attack* in recording on the ball events. In other words, it is much easier to understand or categorise what the attacking team is doing than what the defending team is doing. Clearly in the case of *attacking against a set defence*, the natural interpretation is that these situations are characterised by what the *defence* is doing rather than by what the *attack* is doing. However, for our purposes we had to reverse that interpretation and try and answer *what in the events of an attacking team indicate that they are attacking against a set defence.*

The concepts that we've chosen for our definition are:

- **Duration**: The defensive team needs time to transition into their set defensive shape, and therefore there should be a threshold of duration to consider an event in a possession to be occurring against a set defence.
- **Pinballing**: Despite gaining considerable attention lately, the idea of a *possessions framewor* is a difficult one to execute properly in such a dynamic sport as football. Periods where teams struggle over possession of the ball, for example a string of half-clearances or uncontrolled touches from both teams (which we've referred to as *pinballing*), are hard to include in a possessions framework satisfactorily. For our purposes, we've decided to only consider events in which there is no recognisable *pinballing* for a window prior to the event.
- **Verticality**: Referring to the speed of meters advanced along the axis towards the opponent goal, an *attack against a set defence* can't be progressing too quickly vertically, as this would imply either a full-blown attack during the transition phase, in which case a defence is unlikely to have chance to get set; or an extremely vertical ball against a deep defence which would in any case indicate a divergence from probing against a set defence.
- **Height**: Finally, it also clear that there should be a criteria referring to how high up a pitch the events are taking place (as in, how advanced into the opposition's half). An *attack against a set defence* can't be taking place too far back as that might indicate a team in possession but pinned back by a high block or high press.

How we came to the specifics of the parameters to determine the above criteria will be discussed below.

2.1 Defining the *Duration* Criterion

The main idea that guides how we chose a threshold for the duration of a possession for an opponent's defence to be set is that, *once that has happened, simply letting the time flow should not increase the scoring odds*:

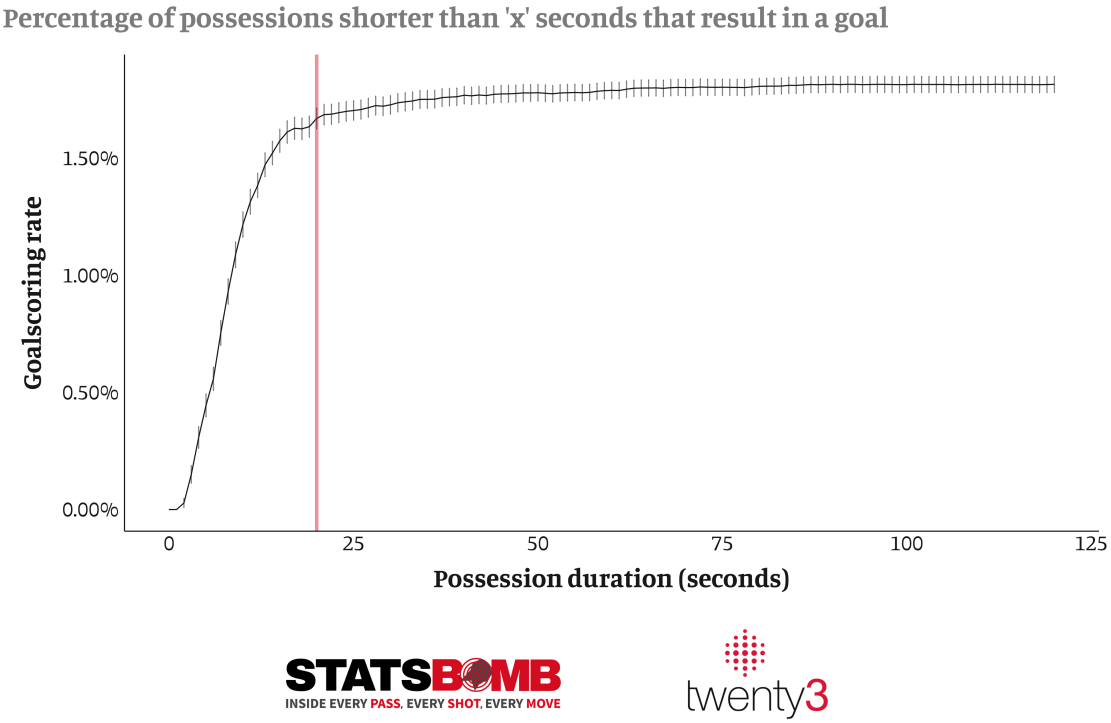
- It is relatively obvious that, for a possession to result in a goal in less than 1, 2, 3... seconds, it needs to be either a set-piece action or a ball recovery high up the pitch that is followed by a quick shot, and we should see very few of them.

- If we increase the time interval we are looking at by a bit, we would find counterattacks, and these, by definition, are not attacks against a set defence, but we should see that, for these, the odds of them resulting in a goal should be relatively high.
- Once a team has had possession for a certain amount of time, and their opponents have transitioned into their defensive shape, the circumstances of the attack don't change with extra time elapsing, and as such the likelihood of these attacks ending in a shot or goal stabilises with time. The starting point for this *plateau* is the threshold we are looking for.

To search for this *plateau*, we need to empirically examine the relationship between possession duration and "results" of attacks. In preparation for this search, we looked only into possessions that have **an attacking action in the final third** for them to be considered as *attacks* (thus filtering out possessions which are for example short possessions of a defending team recovering the ball and then clearing it which would skew our study). Additionally, we also filtered out those where **the team which is considered as being *not* in possession of the ball does more than 20% of on the ball events**[1], as we deem these to be *pinballing* and therefore inserting noise into the effects of possession duration on the success of an attack. Finally, possessions tagged as "From Counter", "From Corner Kick" or "From Free Kick" and with less than five on the ball events are also excluded.

The following chart suggests that **20 seconds** is a reasonable threshold for this issue.

Fig. 1



[1]Here, as throughout the rest of paper, we refer to the ball events as those events tagged as "Pass", "Shot", "Dribble", "Clearance", "Carry", "Interception", "Miscontrol", "Ball Recovery" or "Block".

Throughout the rest of the document, we'll refer to possessions which meet the criteria above and are at least 20 seconds long as **long and well-established attacking possessions**.

2.2 Defining the *Pinballing* Criterion

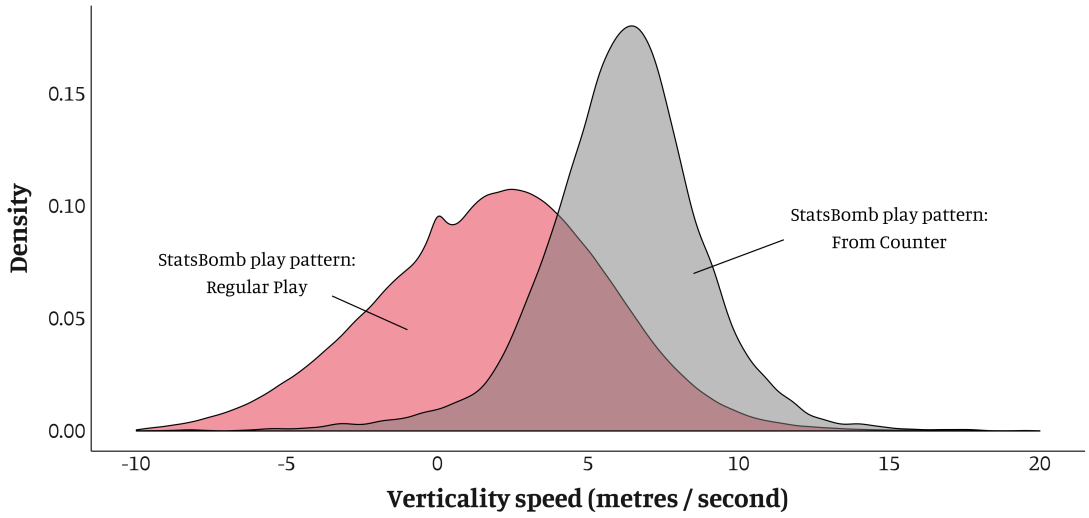
As we have just discussed, we do not want to have possessions (in terms of StatsBomb's schema) where more than 20% of the ball actions are carried out by the not-in-possession team, as we deem these possessions to be too divided and not "natural" possessions and would only insert noise into our sample. Naturally, as we are defining the concept of attacking a set defence *locally*, i.e., event by event, it makes sense to take a similar approach to look only at events such that the possession is not divided in their surroundings. Therefore, we decided to look only at events such that, over the last 20 seconds, no more than 20% of the on-ball actions are carried out by the not-in-possession team.

2.3 Defining the *Verticality* Criterion

This part of the process was fairly simple. Verticality, or vertical speed, needed a threshold boundary as attacks that progress up the pitch at a high speed are unlikely to come against a set defence. Either the attacking team are able to attack at such speed because the defence is not set in the first place, or the speed of attack disrupts the defence out of their set defensive structure.

Fig. 2

Density distribution of the three-event rolling verticality of regular play and counter attacks



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

There's an old adage in football that "the ball travels quicker than the player", a reminder to young players learning the sport that passes travel quicker than they're able to run. Because of the potential speed

of passes, it is necessary to take the progress of numerous events in a row to find the *speed of attack*, rather than taking the speed of single events on their own. We used a three-event rolling window for ball-movement actions (i.e. passes and carries), and calculated the speed that the ball was moved along the long axis of the pitch, comparing actions marked as "Regular Play" and "From Counter" in StatsBomb's play pattern.

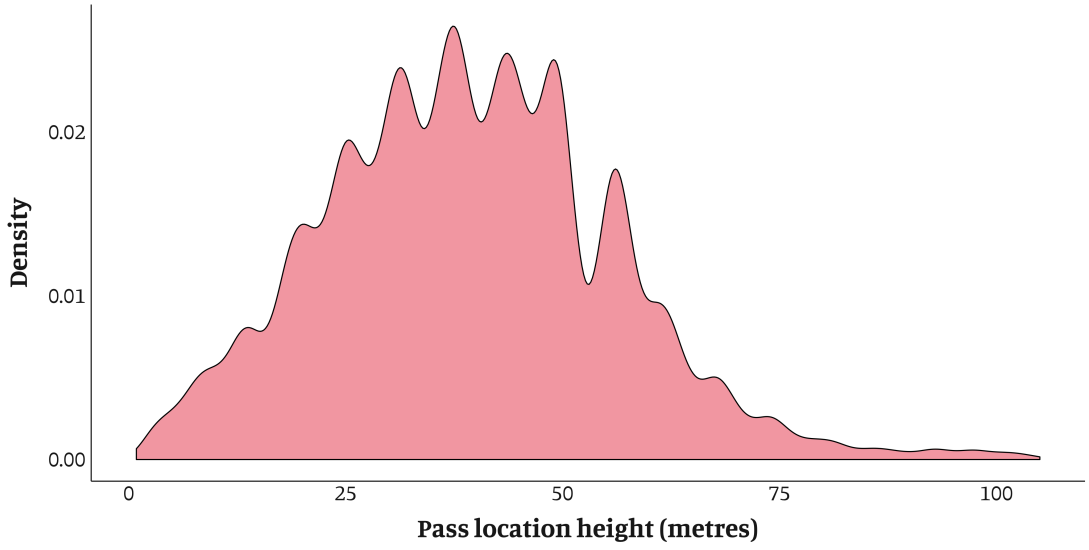
Given that the reason for having a verticality criteria was to rule out quick attacks during a transition phase, it made sense to choose a conservative speed. The vertical speed that we settled on was 4.375 metres per second, which is around the 75th percentile speed in regular play possessions (with outliers of anything above 30 metres per second removed).[2]

2.4 Defining the *Height* Criterion

Conceptually, a criterion for **height** (referring to how close to the opposition touchline a team is) looks to filter out phases of play in which the team in possession of the ball is pinned back in their own half by a high-press, so that we can focus on attacks where the opposition has already backed off to defend in their own half. From a football perspective, we argue that a good indicator for the minimum height a team should be at when the opposition has regressed into a set defence is given by the empirical position in which the centre-backs are making passes in **long well-established possessions** as they were defined in Section 2.1. The intuitive explanation is simple: when a team is probing against a set defence, centre-backs usually form the deepest line of the team and are a usual outlet backwards to keep control of the ball.

Fig. 3

Distribution in the height of centre-back pass start locations



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

[2]We had originally done these investigations using x-coordinates to measure height of the pitch and, therefore, the vertical speed. The vertical speed we settled on, after checking it against some video clips, was 5 x-coordinates per second, which corresponds to 4.375 metres per second, assuming a 105-metre long field of play in relation to StatsBomb's 120-long x-coordinates.

Having seen the distribution of the height of passes from centre-backs, we decided that it made most sense to keep our threshold relatively high, but that we wanted to test some different options. We tested several parameters and watched video of possessions that matched some of these but not others, and came to the conclusion that a height — which turned out to be around the 60th percentile — of 43.75 metres was a good one to use.[3]

This depended partly on the conception of playing against a set defence that we already had in our minds. If we had wanted to look at possessions specifically facing what is termed a 'low block' (when a defending team sits no higher than around 40 metres from their own goal) then we may have chosen a higher value for this criteria.

2.5 Discussion and Example from the Definition

The criteria outlined above allow us to algorithmically classify an event as either True or False: it will be True if it fulfils **all** of the criteria simultaneously, False otherwise.

An important thing to highlight from the definitions above is that we don't classify entire possessions within the possessions framework as occurring against a set defence, but rather individual events or strings of events within those possessions. It is entirely possible for us to consider some events within a single possession as occurring against a set defence, and others within that same possession as not occurring against a set defence. Nevertheless, this doesn't mean that we won't concern ourselves with False events. If, for example, a team has a series of passes which are marked as True where the opposition is set and they're patiently "probing", and then suddenly verticalise play with a through ball into a runner in the box who then shoots and scores, then potentially because of the sudden verticalisation the through ball and shot events are marked as False, but we are still interested in how they went from a True state into a goalscoring opportunity.

Let's step away from the abstract definition through a concrete example of what this algorithm produces on actual string of events in the data.

The given example comes from the game between Crystal Palace and Liverpool in the 2018/19 season at Selhurst Park. Liverpool begin the possession in a probing state knocking the ball around the back as Crystal Palace attempt to implement a press. At this point the speed of which they are circulating the ball is slow enough to imply that they are facing a set defence. However, the height in which they are making their passes is not far enough up the pitch and indicates that rather than facing a set defence they are instead facing a high press from Crystal Palace. Two example events from this first phase our shown below.

Event Type	Possession	Minute	Second	X	Y	Height	Verticality	Team
Carry	62	34	52	30	14	False	True	Liverpool
Pass	62	34	56	33	11	False	True	Liverpool
Carry	62	34	58	24	30	False	True	Liverpool
Pass	62	34	59	24	30	False	True	Liverpool
Carry	62	35	1	29	54	False	True	Liverpool
Pass	62	35	2	29	54	False	True	Liverpool
Carry	62	35	3	45	76	False	True	Liverpool
Pass	62	35	7	52	75	False	True	Liverpool

[3] We had originally done these investigations using x-coordinates to measure height up the pitch and had used round numbers in the study. The x-coordinate height that we settled on was 50, which corresponds to 43.75 metres up the pitch, assuming a 105-metre long field of play in relation to StatsBomb's 120-long x-coordinates.

In this phase their play is too deep and thus does not trigger the necessary height parameter, causing it to remain False. Liverpool continue to pass in this manner, slowly working their way up the pitch, eventually reaching a phase where both the height and the verticality conditions are met.

Event Type	Possession	Minute	Second	X	Y	Height	Verticality	Team
Carry	62	35	36	64	62	True	True	Liverpool
Pass	62	35	37	64	62	True	True	Liverpool
Carry	62	35	39	63	37	True	True	Liverpool
Pass	62	35	40	65	37	True	True	Liverpool
Carry	62	35	41	72	23	True	True	Liverpool
Pass	62	35	42	72	20	True	True	Liverpool
Carry	62	35	43	86	10	True	True	Liverpool

In this phase Liverpool are still manoeuvring the ball from side to side, at a speed slow enough to keep the verticality criteria True, as they look for an opening. This is similar to the first phase, however these passes are far higher up the pitch, thus meeting the criteria for height and inferring that they are now facing a set defence.

It is important to recognise that, inevitably, the complexity of strictly labelling situations in such a fluid sport as football compounded by the attacking bias of the on-the-ball events we mentioned above implies that the algorithm won't be perfect and will have both false positives and false negatives when compared to human judgement of these game situations. Nevertheless, throughout the rest of the document we hope to appease doubters by establishing that the situations we've marked as True have significant signal of interest, and have natural and robust implications for practitioners.

3 Empirical Results of *Attacks Against a Set Defence*

Section 2 has left us with a set of criteria with which we can algorithmically define events which we consider to occur against a set defence (marked as True). In this section, we'll focus on looking at the observed outcomes of these situations, and contrast it with the situations which we've marked as False.

3.1 Observed Outcomes: A Simple Frequentist Outlook

A first step towards understanding the importance of the *type of possession* in our study is to simply look at the most frequent results of such possession types, i.e., how many of them result in a shot or a goal, and what is the total xG accumulated by each possession type. Lets start by comparing these magnitudes for three types of possessions:

- **Attacks against set defences:** Possessions that contain an event marked as True as described in Section 2.5.
- **Other long and well-established attacking possessions:** As in 2.1, recall that this refers to possessions that
  - are longer than 20 seconds,
  - at least 80% of the ball actions are by the in-possession-team,
  - and that are not tagged as "From Counter", "From Corner Kick" or, if they are tagged as "From Free Kick", the in-possession-team has had at least five ball actions.

  The fact that a possession like this is not flagged as an attack against a set defence must then be due to the fact that for all of its events, either the verticality or the height criteria fail, i.e., at all times we should see that either the possession is too vertical or the actions are happening too far back in the in-possession-team's own half of the pitch. Therefore, these involve both possessions that may not be considered to have a really offensive intention as well as others that, by their *directness*, may indicate that the not-in-possession-team is not well set to defend. An example might be a team pinned back by a high-press early in the possession, which eventually breaks through and manages to attack a defence which in nature almost resembles that faced by counter-attacks.
- **Counter-attacks:** Possessions marked as 'From Counter' in Statsbomb's play pattern field in the data.

The results are tabulated below:

	Avg. goals by possession	Avg. xG by possession	Avg. shots by possession
Attacks against set defences	0.020	0.018	0.185
Other long and well-established attacking possessions	0.012	0.010	0.069
Counter-attacks	0.048	0.048	0.312

We see quite clearly that achieving outcomes when attacking against a set defence is harder than when counter-attacking, both by looking at the goal-scoring rate but also at the ratio between goals scored and shots: 10.81% of shots result in a goal when attacking against a set defence, whereas this increases to 15.38% if we look at counter-attacks. The fact that there's better outcomes when attacking against set defences than in other long attacking possessions may very well be explained that the proxies we have used to construct those (*not too vertical*, *not too fast*) implies a certain degree of *calm* that leads to smaller chances of losing the ball (in fact, 70.35% of attacks against set defences end with a ball loss, whereas this percentage increases to 75.78% when we look at other long attacking possessions which might represent for example teams pinned back by a high-press which never manage to mount an attack despite having possession).

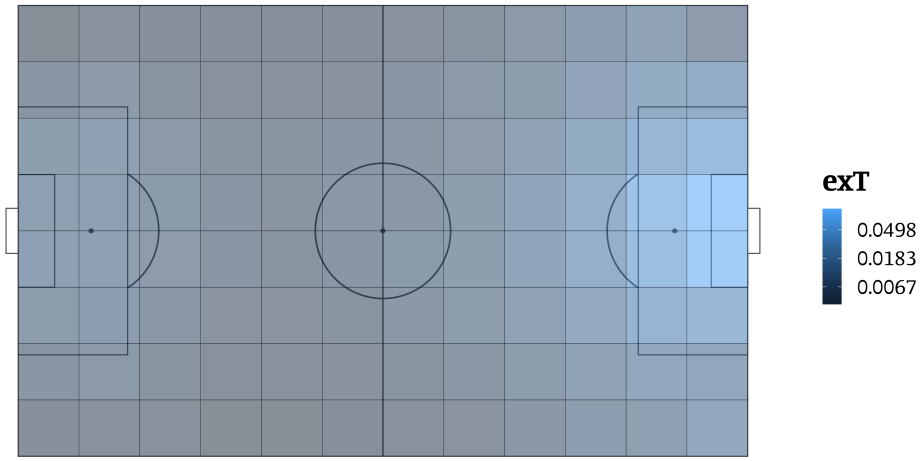
3.2 Empirical xThreat:

Let's take a step further from the frequentist approach, and try and explain in more detail where the difference in observable outcomes is coming from between True and False runs of play. A natural place to go looking for this additional detail is to understand from which zones of the pitch teams are generating the most danger depending on whether they're attacking in a True or False situation. Along this line of investigation, Karun Singh's *expected threat* (xT) approach [2] has gathered attention for its simplicity, interpretability and applicability. His basic idea is that from any given zone on the pitch, an attacking team can either shoot (with a certain xG value), or move to a different zone according to a transition probability matrix. A recursive formula allows to solve for the xT value of each zone of the pitch. His work shows that the recursive formula stabilises towards the final values quite quickly, after only 5 or 6 iterations. The interpretation of this is that it suffices mostly to consider the likelihood of scoring within the next 5 or 6 moves (with moves being either passes, carries from one zone to another, or shots) to have an approximate appraisal of the threat associated with being in a given zone.

With this line of thinking, we computed a simple metric which mimics xT which we've labeled as *empirical expected threat* (exT), which for each zone of the pitch is defined by the empirical likelihood of scoring within the next 5 moves. As *"zones"* we've used a 12×8 grid of the pitch. The visualisation below illustrates the different levels of exT associated with the different zones of the pitch.

Fig. 4

Empirical xT of all events



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

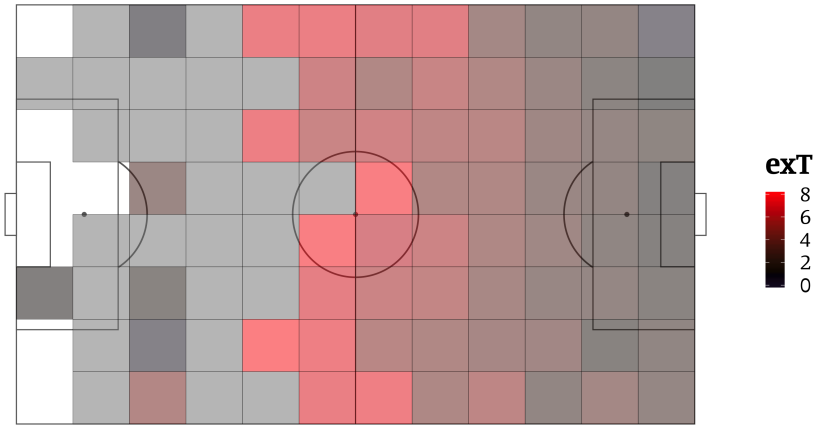
twenty3

The simplicity of exT (it's basically a frequentist metric) allows us to compute it for any subset of events. In particular, we can compute exT solely for True events, and compare the results with the exT for the phases of play examined in Section 3.1.

Compared to events in counter-attacks, the counters clearly have the upper-hand, with the relative exT value of counter-attacking actions as much as eight times greater than when facing a set defence. The scale is based on the attacking half of the pitch, with the defensive half in many of the following examples being victims of noise (our *height* criterion ensures that not too many actions against this far back are marked as True so the sample size for these locations is small).

Fig. 5

Relative exT values of events in counter-attacks vs events against a set defence



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

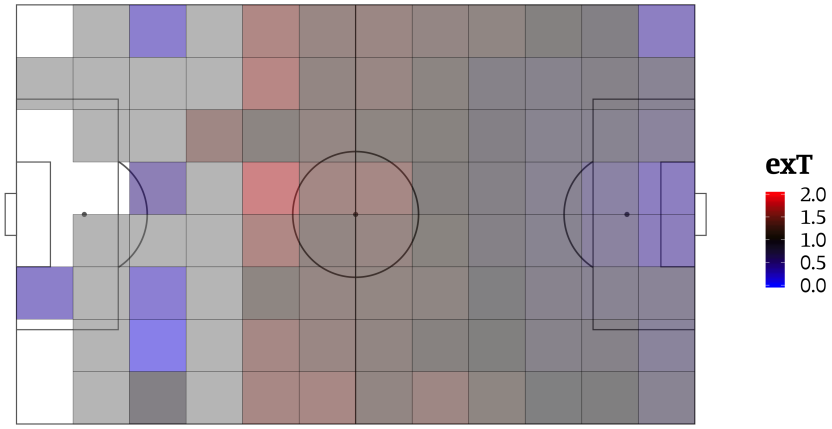
twenty3

However, the increased threat of counter-attacks diminishes the closer to goal that we get. Intuitively, this makes sense: it would be more effective to attack quickly when further away from goal, but the relative advantage of that diminishes when there is less space in which to attack into, and when defences will likely be more compact.

In comparing events that come against a set defence to those that are in established possessions that aren't against a set defence (i.e., that fail one or more of our other parameters), playing against a set defence appears to be slightly more threatening, but only when closer to goal.

Fig. 6

Relative exT values of events in long, 'non-asd' possessions vs events against a set defence



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

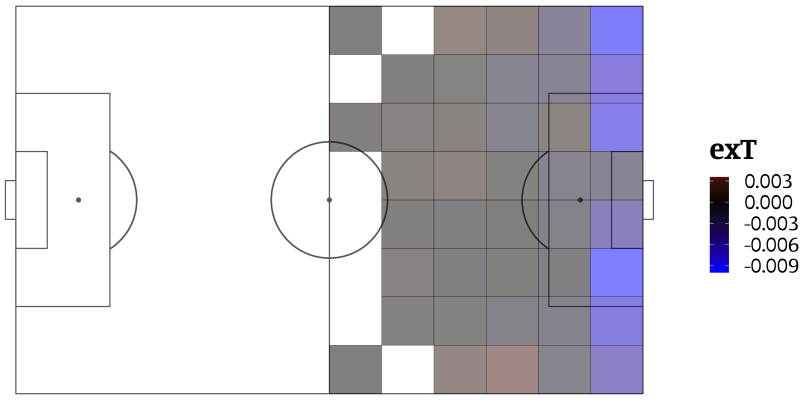
Potentially, there is an advantage to being closer to goal when playing against a set defence than in established possessions that may be less settled: this may be because teams are in more of an attacking structure and better able to construct good attacks, even though the defence is likely in a set defensive structure of their own.

Notice that the two zones that cover the six-yard box are coloured in a way that indicates that playing against a set defence creates more of a threat in these areas. To investigate further, we can look at where teams get into these two zones from.

Below is a visualisation which looks at what zones make successful passes into those two "six-yard box zones". Once again, we are comparing events that come from long possessions that aren't necessarily against a set defence against events which fit all of our parameters of playing against a set defence.

Fig. 7

Location of successful passes into the six-yard box zones; absolute difference in proportion between events in long, 'non-asd' possessions and those against a set defence



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

The dominant zones where proportions of successful passes in the six-yard zones come from are closer to the byline against a set defence than when not against a set defence. The zones where a larger proportion come from not against a set defence than against a set defence are more like deeper crossing zones, which are likely to lead to lower value chances than those resulting from passes from the byline.

The success rate of passes from these zones in the final row of the pitch into the six-yard zone also plays a factor. The success rate of zones (11, 0-2 and 5-7) to zones (11, 3-4) is 21.59% in long possessions that aren't against a set defence while it is 24.39% in our parameters of being against a set defence.

It appears that events that match our parameters of playing against a set defence include passes completed at a higher rate in these dangerous areas of the pitch, though, which may be due to the advantages of having a more established attacking structure. The evidence seems to point towards the fact that once you're near the byline against a set defence, you'll be in a position to pass successfully into the box (executing a "cutback") more frequently than in other game situations, and as such the relative "disadvantage" in threat from other locations of the pitch when against a set defence starts to diminish caused by the presence of this dangerous option in attack.

Whether this result is directly actionable is a different question, as teams will already know that getting to the byline will enable them to carve out dangerous chances. In the next section we look to isolate factors leading to success even further.

4 *How to Break Down a Set Defence:* Data-Driven Recommendations on Factors of Success when Attacking a Set Defence

Up to this point we've focused on defining, describing, and evaluating the results of the attacks we've considered to be happening against a set defence. In this section we dial up our ambitions and try to go deeper into the in and outs of attacks against a set defence, and the factors that determine whether they're successful or not. In trying to answer this question, we've separated it into two different aspects:

- **True Events:** In Section 4.1, we attempt a simple feature engineering approach to model the likelihood of a True event to lead to a goal in the near future. From the decision function we can try and retrieve the individual signal of each feature in an attempt to understand what about a True event can indicate higher likelihood of success.
- **False Events:** As we saw in 2.5, for every possession with True events, it will in all likelihood also have string of events where the team does something different and transitions into a False state for a few consecutive events, which we'll refer to as **False Blocks**. Arguably, the moments in which a team breaks away from "probing" and attempts something are also key moments that determine the outcome of the attack. Section 4.2 explores the insight to be found here in a completely independent approach to what we showcase in 4.1, by clustering and classifying different types of False blocks into different "choices" the team made, and evaluating the empirical reward.

Finally, in Section 4.3 we also turn our attention to the trade-off of reward vs risk in situations where a team is attacking a set defence, by evaluating what can happen when they lose the ball and their opponents then take their turn at attacking, and relate the results to the factors of reward we've studied in 4.1 and 4.2.

4.1 Feature Engineering to Predict the Reward of True Events

Conceptually, the first question to ask ourselves is "what features about True events can be used to model the likely outcome of these attacks?". Clearly, a True event is a single snapshot in time, and if we restrict ourselves to these snapshots as such then we'd be basically constructing an *expected goals* models for shots occurring against a set defence. Instead, in the same spirit as 3.2, a better proxy for the "outcome of an event" is whether or not a goal will be scored within the next 5 moves (with moves specifically being either passes or carries). Marking each True event as either 1 (there is a goal in the next five moves) or 0 (there isn't a goal in the next five moves) gives us a label to model against. Section 4.1.1 specifies the regressors that we've built ahead of the actual modelling.

4.1.1 Feature Selection

There are a few features, like the distance to goal or whether or not the event was under pressure, which are clear candidates for regressors to whether or not the attack will lead to a goal. Nevertheless, following a similar line of thinking to the outcome being "within a couple of moves", in constructing features associated to each True event it also makes sense to look a bit into the past to understand what has been happening in the lead up to each event in the sample. The following features were selected and constructed for our sample:

- **Under Pressure**: This feature is directly available in StatsBomb's data, marking whether the player was under pressure from an opponent when performing the action or not.

- **Distance to Goal**: Number of meters from opponent's goal.
- **Rolling Width**: On a 3-event rolling basis (passes and carries only), how many meters of the pitch has the attacking team covered horizontally.
- **Rolling Speed**: On a 3-event rolling basis (passes and carries only), how many meters per second has the ball travelled.
- **Rolling Time Between Consecutive Passes**: On a 3-pass rolling basis, how many seconds have there been on average between consecutive passes. Notice that the higher this number, the more seconds have elapsed on average between consecutive passes, meaning passes are coming about more *slowly*.

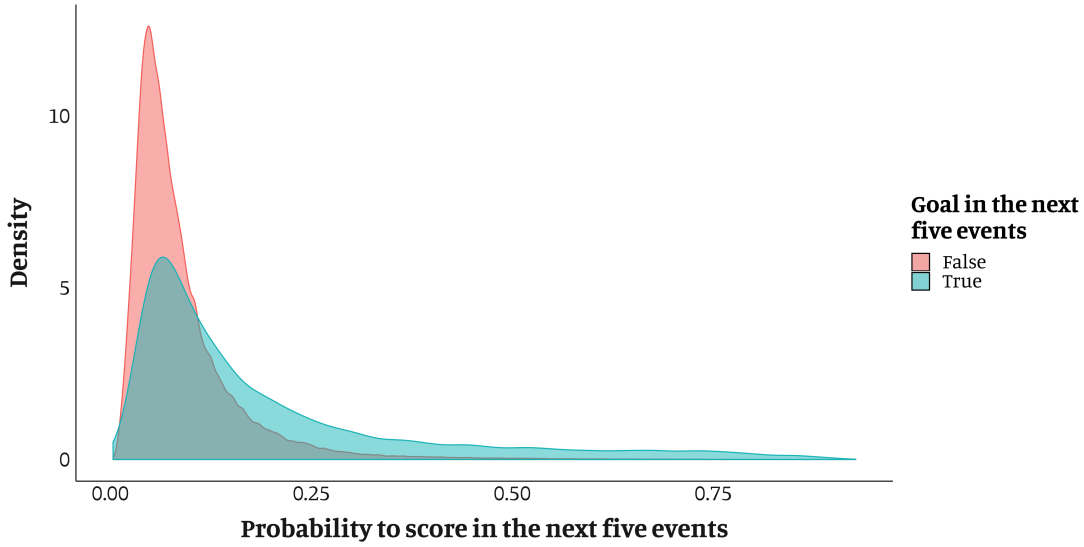
4.1.2 Modelling

Predicting the outcome of events from the features selected is a challenging prospect. On the one hand, the sample is heavily unbalanced. Specifically, out of 994,725 True events in our sample, only 12,832 have a goal scored within the next five moves. In addition to this problem, the regressors are clearly not independent, and a simple regression model like a logistic regression won't perform well. In light of these circumstances, we've chosen to first apply a resampling technique to the sample, which we then feed forward to an ensemble random-forest based gradient boosting algorithm (using the well known *XGBOOST* library in Python).

Evaluating the performance of the trained model is difficult precisely because of the unbalanced sample. However, there are positive signs that the model is learning valuable structure in the sample. On the one hand, in the resampled test sample (where roughly 50% of observations are marked as having a goal within the next five moves), the model's classification is 95.7% accurate. Additionally, when we pass the actual original sample through the model and assign the "goal probability" value to each sample (as opposed to the binary 0-1 classification), we can see that the probability distribution for events which in reality resulted in a goal within the next five moves is markedly skewed towards higher probabilities than events that didn't.

Fig. 8

Density plot for the predicted probability to score in the next five events for each class



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

4.1.3 Evaluating Feature Impact on Goal Probability

The modelling section has left us with a degree of confidence that we have a model which captures a certain degree of structure in our sample. However, we now face the fresh challenge that as opposed to simple models, forest-based models are powerful in learning but don't lend themselves to simple interpretation of individual feature importance or impact on the regression, which is what we're after for practical advice on these game situations. However, there are creative techniques to circumvent these difficulties and manage to glean insight into the impact of individual features under certain circumstances. For this project we've attempted an innovative approach to isolate the individual effect of features, which is outlined below:

1. First we carry out a K-Means clustering process on the feature space. The reason for this is that forest-based models tend to learn structure in a "local" way, i.e. they can predict new observations well when they're in the vicinity of a dense set of training samples. By clustering the feature space, we hope that within each cluster our model has clearly learned in that area of the feature space.

2. Using the partition of the sample space generated by the clustering, within each cluster we fix its centroid, and taking turns vary each individual feature from its 10^{th} percentile through to its 90^{th} percentile *within that cluster specifically*, and pass this synthetic sample to the model to record its "goal probability".

3. For each cluster, by plotting the feature's value versus the resulting probability prediction we can understand how, *within that cluster specifically*, the feature's value affects the likelihood of the event leading to a goal within the next five moves. In essence, it lets us understand feature importance at a

very local level. Because of the potentially highly non-linear nature of random forests' decision functions, this technique of "fixing all variables bar one which we vary" can potentially be very misleading if done globally. Hopefully, through the local partition of the sample through clustering we've managed to circumvent this pitfall.

4. Note that the *under pressure* axis wasn't considered as part of the feature space when clustering. When creating synthetic intra-cluster samples, we simply created two, one with each True/False *under pressure*.

The table below specifies the centroids of the different clusters:

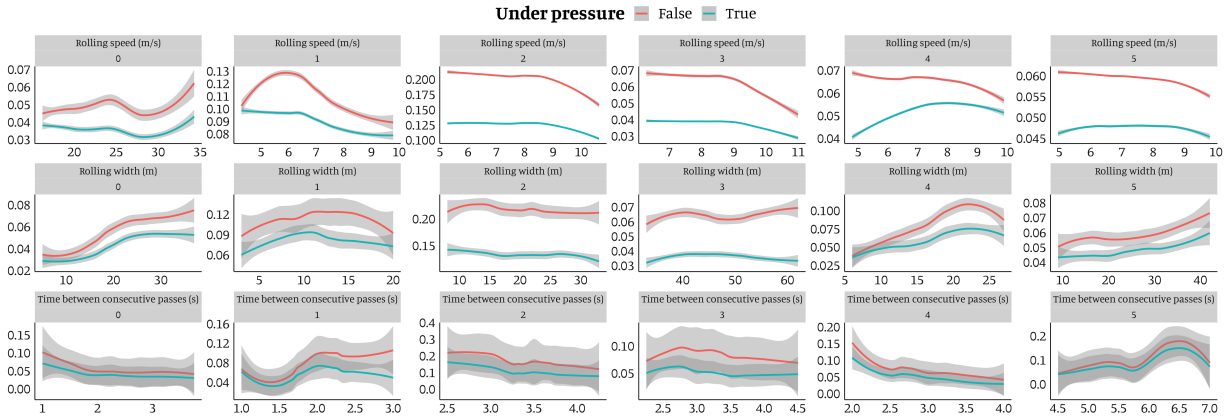
Cluster Centroids

	Distance to Goal	Rolling Width	Rolling Speed	Time Between Consecutive Passes
Cluster 0	44 m	18 m	22 m/s	2.4 s
Cluster 1	39 m	10 m	6.7 m/s	2.1 s
Cluster 2	27 m	19 m	7.7 m/s	3.3 s
Cluster 3	47 m	45 m	8.4 m/s	3.2 s
Cluster 4	55 m	16 m	7.1 m/s	2.8 s
Cluster 5	43 m	23 m	7.3 m/s	5.7 s

With the above cluster centroids in mind, the figure below showcases for each cluster and feature, how the model's "goal within the next five moves probability prediction" evolves as we vary each feature while the rest remain fixed at the centroid. As we mentioned before, *under pressure* is binary and as such we simply create two synthetic samples each time for its values. Also, we don't vary *distance to goal* because it's quite clearly increasing: the closer you are to goal the more likely you are to score within the next five moves. Therefore we are showing the results of varying intra-cluster *rolling speed*, *rolling width* and *rolling time between consecutive passes*.

Fig. 9

Predicted probability to score in within the next five events as we vary the classifier parameters in each cluster



There are several interesting conclusions to draw from the results.

Clusters 0 and 5 exhibit a tendency to be more dangerous (as in, more likely to result in a goal within the next five moves) as *rolling width* increases. Both clusters are centred at a very similar distance to goal, 43 and 44 meters respectively; and are also centred at a similar *rolling width*. What differentiates both clusters is related to the speed of movements: Cluster 0 is the cluster with the fastest *rolling speed*, whilst Cluster 5 is much slower and is the cluster characterised by being markedly the one with the most amount of seconds between consecutive passes. Whilst both share the same tendency in *rolling width*, they have contrary tendencies in *rolling speed* and *time between consecutive passes*, meaning that within each cluster, the already dominant strategy is advisable: in Cluster 5, slowness is preferable; in Cluster 0, fastness is preferable; but in both clusters, the wider you make play, the more dangerous you'll be.

Cluster 2 is the cluster whose centroid is closest to the opponent's goal, and accordingly is the cluster with the highest probability prediction on average. Its interesting that the prediction is fairly stable as we vary the features, with the only real signal being a drop-off in the probability once speed increases past 8 m/s, which should be interpreted as there being a threshold when close to goal above which increased speed of play decreases the likelihood of you scoring.

Cluster 1 is interesting to analyse compared to the rest: its centroid has the lowest *rolling width*, the lowest *rolling speed* and the least *rolling time between consecutive passes*. In football terms, it's a cluster of samples characterised by taking place in a narrow strip of the pitch, with quick exchange of passes that don't cover much ground over time, potentially one-two's in close proximity. It's interesting in that it's the only cluster in which features exhibit clear optimums and minimums. In *rolling time between passes*, the probability reaches a clear minimum around 1.5 seconds between passes. For *rolling width*, the prediction reaches an optimum between 10 and 15 m wide. Finally, the case of *rolling speed* is fascinating in that when **not under pressure**, there is a clear optimum around 6 m/s. However, when **under pressure**, this peaking between 4 and 6 m/s speed is not observed, meaning that that increase in speed is only beneficial if not exchanging passes under pressure.

This final point on pressure is also interesting to highlight. Satisfactorily, across the board we can see that the model understands that being **not under pressure** has a higher chance of leading to a goal within five moves than being **under pressure**, as is intuitively expected. It is interesting however to analyse the relative difference between the True/False *under pressure* states. The most interesting case is that of *rolling speed* in Cluster 4, because the True/False *under pressure* curves experience opposite tendencies: between 5 and 8 m/s: whilst the probability *decreases* in a True case, it *increases* in the False case. Observing the Cluster 4's centroid is the furthest away from goal (55 m), in a footballing sense, this is saying that when against a set defence that far from goal, you should slow down if not under pressure, and speed up if under pressure. Potentially, if you're under pressure and you speed up play that far from goal, you're dragging the markers who are pressuring you out of position and creating space to string together an attack further upfield.

It's important to close out this subsection with a word of caution: interpreting these results is hard, and we may be drawing conclusions that are noise in the model rather than corresponding to true structure. Nevertheless, the random-forest based model clearly captured *some* signal in the data, and the technique shown above makes an attempt to isolate the effects of features from that model. Even readers who doubt the validity of this technique in particular, should appreciate the opportunity for other model-agnostic methods used on well-performing black box models such as these to try and find practical results.

4.2 The False blocks: classification and reward

The discussion above has mainly looked at the concept of "attack against a set defence" and the outcomes from those attacks (or, more precisely, from possessions that contain those attacks). However, as visual inspection suggests (see Section 2.5), the goal-scoring chances and the moments of *high* risk of losing the ball may happen inside blocks of consecutive False events (False blocks), i.e., moments inside those long attacking possessions in which there is either a change in pace, height, verticality or *pinballing* that takes the possession out of a True state.

We've attempted to understand and classify these "False blocks" as follows:

- Looking only at False blocks that happen in a possession that contain a True state, compute the following features for each of them:
  - The reason why the block is False (i.e., if it contains an event tagged as *pinballing*, *lack of height*, or *high verticality*).
  - Number of players involved.
  - Number of ball events that encompass the block.
  - Where the block is located (determined by the centroid of the locations of events that form the block).
  - The *spread* of the block.
  - Block duration.
  - Speed of ball circulation (computed as the average speed of the last four passes and carries) in the event that leads to the block.
- Run several iterations of a clustering algorithm to isolate groups of False blocks of similar nature, and assign a certain *type* to all blocks belonging to the same group.
- Get an understanding of the risk and reward associated to blocks of every type by computing:
  - How the odds of scoring increase/decrease for possessions that contain blocks of each type with respect to possessions that don't contain blocks of that type.
  - How the chance of getting a shot increases for possessions that contain blocks of each type with respect to possessions that don't contain blocks of that type.

4.2.1 Clustering of False blocks

Using several iterations of a K-Means clustering algorithm on this 11-dimensional feature space, the following significant clusterings were discovered. As we could have guessed, the reasons why the blocks are False are important factors in defining clusters. However, their combination with other features such as location and speed of events yield descriptions that are easily interpretable in footballing terms[4]:

- **Hard reset:** The block contains an event tagged as *lack of height*, its centroid is less than 35 meters away from the in-possession-team's goal-line and the number of on the ball events is less than half of the duration of the block, in seconds.

[4]The definition of each of the clusters should be understood as "All blocks that do not belong to any of the clusters above and..."

- **Mild reset:** The block contains an event tagged as *lack of height*, its centroid is in the in-possession-team's own half and the number of on the ball events is less than the duration of the block, in seconds.
- **Pinballing:** The block contains an event tagged as *pinballing*, i.e. neither team had a fully controlled possession of the ball for events in that block.
- **Fast wing-play:** The block centroid is less than 15 meters away from either of the wings and the speed of ball circulation in the event that leads to the block is greater than 10 meters per second.
- **Fast central combinative play:** The block centroid is more than 15 metres away from both of the wings, the speed of ball circulation in the event that leads to the block is greater than 10 meters per second, and there are at least two players involved in the block.
- **Fast central individual play:** The block centroid is more than 15 metres away from both of the wings, the speed of ball circulation in the event that leads to the block is greater than 10 meters per second, and only a player is involved in the block.
- **Other**

The following table shows the number of False blocks in each cluster:

	Number of False blocks
Hard reset	124,637
Pinballing	143,133
Fast central individual play	95,253
Fast central combinative play	10,681
Other	252,242
Mild reset	60,038
Fast wing-play	97,936

It is relatively clear that there should be significant differences in the risk and reward teams take when they find themselves into each of the blocks above, in terms of how likely they are to spring a successful attack, or to lose the ball right afterwards. Lets evaluate the "reward" associated to each block.

4.2.2 Empirical Reward of False Blocks Clusters

Reward metrics For every cluster of False blocks, we define the following metrics:

$$\text{Goals lift} = \frac{\text{Average goals by possession when the possession contains a block of that type}}{\text{Average goals by possession when the possession does not contain a block of that type}};$$

$$\text{Shots lift} = \frac{\text{Average shots by possession when the possession contains a block of that type}}{\text{Average shots by possession when the possession does not contain a block of that type}}.$$

The results broken down by cluster are:

	Goals lift	Shots lift
Hard reset	2.229	2.590
Pinballing	1.490	1.291
Fast central individual play	1.354	1.547
Fast central combinative play	1.178	1.120
Other	0.991	1.286
Mild reset	0.933	1.285
Fast wing-play	0.745	0.925

The results clearly point towards *hard reset* as the highest reward option. Intuitively, this makes some sense: "resetting" the possession is intended to draw the defending team out of their deep defensive structure and open up space to attack into. It makes some sense that this strategy, coached and used in the professional game, has some quantitative value.

It is also quite interesting that there is such a significant difference between a "hard" reset and the "mild" one. Also, that "pinballing" and "fast central combinative play" yield a greater lift in goals than in shots in contrast with the rest of the block types, which suggests that the shots produced by these are easier to convert (which makes sense, as one can easily imagine both of these to end with central inside-the-box shots).

4.3 Evaluating *Risks* whilst Attacking a Set Defence

Clearly in such a dynamic and fluid sport as football, risk and reward are constantly interplaying with each other: attempting a dribble might create a dangerous opening if successful; but performing it may lead to a ball loss that exposes your team to a counter-attack. In light of this, in this section we shift our attention to the *Risk* dimension of the factors of attacks against a set defence which we've been studying.

4.3.1 Risk of Goal Probability Features

In 4.1 we defined a set of features which we use to predict the likelihood of a goal coming within the next five moves, and how that relates to the individual features themselves. In this section we want to flip the question, and understand how they relate with the risk of the opponent recovering the ball, shooting or scoring.

The approach must vary slightly in that as opposed to the case of reward, an events features should be held accountable for the opponent shooting or scoring *in the possession immediately following the ball loss*, not just 5 events from it. To evaluate *risk* in this way, we compare the average of features in the lead up to losing the ball, and compare that against the proportion of opposition possessions which end in a shot or goal. The figure below plots the results.

Fig. 10

Risk with respect to distance to goal

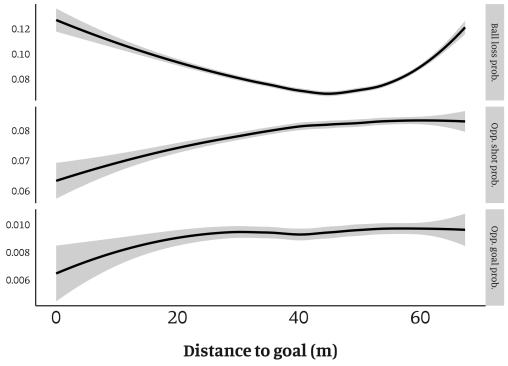


Fig. 11

Risk with respect to rolling width

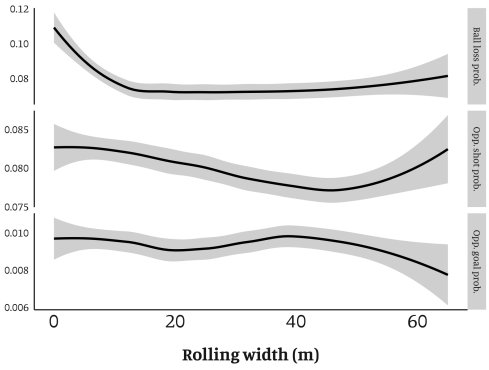


Fig. 12

Risk with respect to rolling speed

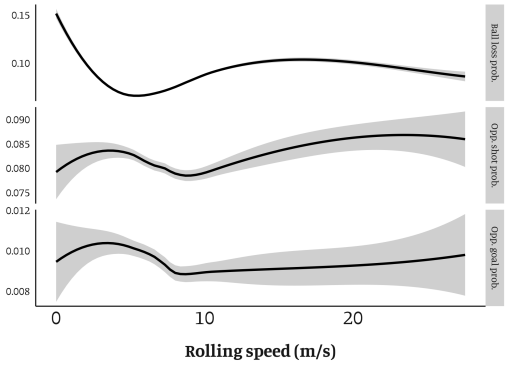
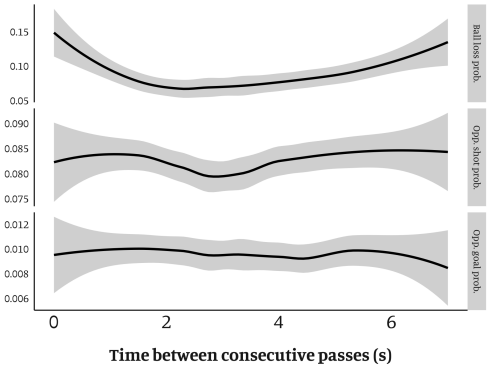


Fig. 13

Risk with respect to rolling time between consecutive passes



STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

twenty3

As in the case of "reward", we're hard pressed to extract concise conclusions, but there are interesting elements in the results. For *distance to goal*, we are clearly picking up the expected result that the further from goal, the lower the probability of losing the ball, but the higher the likelihood of conceding a shot or goal in the event of losing it. Its very interesting to observe that there seems to be an optimum "holding" height when attacking a set defence: around the 40m from the opponent goal mark. Any further than that you increase your likelihood of losing the ball; but likewise, any deeper than that into your own half and you also increase that probability, most likely by triggering a press from the opponent.

The results for *rolling speed* also have a similarly interesting trade-off. Increasing speed up to 6 m/s is clearly decreasing the likelihood of losing the ball, but increasing the likelihood of *if the ball is lost*, then the team will concede a shot or goal. This might point to the fact that increasing speed necessarily requires more men to be committed to moving the ball quickly and showing to receive passes, and they might be less set up to track back in the event of a ball loss. Compounded with the results from Section 4.1.3 which showed that 6 m/s was a local optimum of speed under certain circumstances in terms of probability of scoring a goal in the next five moves, certainly that threshold of speed has interesting properties in these game situations.

4.3.2 Risk of False Blocks

Similarly, in Section 4.2 we defined a classification of what we've referred to as False blocks, and studied the reward associated with their presence. How about the risks associated with them?

For each type of False block, we define:

$$\text{Probability to lose the ball} = \frac{\text{Number of times a block of that type ends with a ball loss}}{\text{Number of times a team gets in a block of that type}},$$

$$\text{Probability to concede a shot} = \frac{\text{Number of times the team concedes a shot in the next possession}}{\text{Number of times a block of that type ends with a ball loss}},$$

$$\text{Probability to concede a goal} = \frac{\text{Number of times the team concedes a goal in the next possession}}{\text{Number of times a block of that type ends with a ball loss}}.$$

Then, we have

	Prob. to lose the ball	Prob. to concede a shot	Prob. to concede a goal
Hard reset	0.113	0.077	0.008
Pinballing	0.203	0.084	0.010
Fast central individual play	0.082	0.092	0.010
Fast central combinative play	0.239	0.082	0.011
Other	0.071	0.086	0.011
Mild reset	0.157	0.089	0.008
Fast wing-play	0.066	0.082	0.009

As one could expect, the resetting options are the safest in terms of conceding chances, and the fact that we don't see small probabilities to lose the ball is due to the fact that, with a significant frequency, these are followed by long balls (which, in return, make the opponent recover the ball in a not dangerous area). Naturally, central play is the riskiest of the options.

5 Conclusions and Future Work

We've presented the question of *how to break down a set defence* end-to-end; from defining a proxy, through to isolating the risk and reward of different factors of the attack. This exercise has left us with several conclusions.

In terms of the actual proxy, evaluating its effectiveness at scale is inviable: attacking a *set defence* is a subjective concept of football and, without tracking data of defensive players, a tenuous concept for event data to represent. Despite this inherent difficulty, we don't believe it should take away validity form the rest of the paper's results. As with other data proxies in football, such as the ubiquitous xG, regardless of whether they proxy an actual subjective concept or not (xG seeks to proxy the danger of a shot), if the heuristics used are natural and there is signal in the structure of the definition, the results are interesting in their own right. In our case, the heuristics we've used are natural for a "slow, probing possession in which

the attacking team is mostly in the opponent half and has established control of on-the-ball events, and has given the defending team enough time to transition backward", and we believe this is enough to claim the results are interesting in their own right.

The definition of "set defence" which we arrived on encompassed, in footballing terms, both a mid block as well as a low block. Further research may wish to concentrate on one of these two defensive systems more specifically, or may wish to compare the two, as different approaches may work better for one over the other. For example, it may be that a "hard reset" in the possession has more of an effect against a low block than a mid block. Different practitioners may also choose to set different definitions or proxies for a set defence for different footballing reasons, and different data sources would provide a range of opportunities for exploration. The most obvious is tracking data, of course, which would mean that a set defence can be defined directly rather than through proxies.

In terms of the actual results of what we defined to be *attacks against a set defence*, we've also found some interesting takeaways. In terms of exT locations, we saw in Section 3.2 that as an attack approaches the opponent's byline, the relative "disadvantage" of *attacks against a set defence* compared to other situations like counter-attacks starts to disappear. As we saw, a possible explanation of this result is the greater proportion of cutbacks from the byline into the box from *attacks against a set defence* as opposed to other situations. Cutbacks with the defenders reversing towards their goal whilst attackers run into space are notoriously dangerous situations; and there seems to be a point that these are crucial routes towards creating chances when attacking a set defence. This line of research in football, the concept of *expected threat* as assigning probabilities to points in time in *processes* of attacks, is clearly rich and full of opportunities. We've presented a very simple frequentist approach, but it'd be very interesting to return to evaluate our proxies with other more sophisticated approaches of this type in future work.

In Section 4.1 we built a black-box model to evaluate the potential reward of different events against a set defence, and carried out an innovative approach to try and isolate the individual effects learned by the black-box. We highlighted a set of possible conclusions to draw from the results, mainly relating to the increased danger in increasing the *rolling width* of the attack, or around the "sweet spots" of speed of play between 5 and 8 m/s, which came up again in Section 4.3.1 indicating a trade-off between potential reward versus risk of losing the ball and conceding chances of your own. However, we'd argue that the main takeaway from this exercise is the potential in attempting model agnostic techniques to analyse results form black box models. Football is clearly a complex sport and modelling or predicting outcomes from features of play undoubtedly needs powerful machinery. However, black box methods draw a wedge between the results they produce and practitioners who need to understand the practical implications of individual features. We believe that more attempts of this kind will deeply benefit the applicability of machine learning in football

Section 4.2 took a different viewpoint in analysing not events marked as True, but rather what we defined as False blocks within possessions that were True as some stage. Through a clustering technique, we've attempted to classify different types of False blocks, and have found that what we've labelled as a "*hard reset*" is the highest reward option. Intuitively it's an interesting result, as a *hard reset* is aimed at drawing the defence out and creating space to spring an attack into. Our results provide evidence that this strategy might be successful. It's interesting that there is such a stark difference in the results between "hard" and "mild" resets, both in reward and risk. The results point towards the maxim: "if you're going to reset play, reset it properly".

Finally, Section 4.3 evaluated the risk associated with the factors we had set up to study their reward. Whilst some interesting results were achieved, the study is lacking in a fundamental aspect: the techniques

to evaluate risk and reward weren't immediately comparable. The reason for this is that whilst in attack we consider reward coming in the near future (next few moves), in *risk* we must consider the outcome of the whole possession upon losing the ball, not just the first few moves of the opponent. Likewise, in risk there is arguably a greater weighting to the actual event of ball-loss. In any case, our framework isn't set up for a directly comparable "*game theoriesque*" question of optimal strategy. This is an interesting direction for future work.

References

[1] Opta Pro, *Blog: Phases of Play - an introduction*, available at `(https://www.optasportspro.com/news-analysis/phases-of-play-an-introduction/)` (2019).

[2] K. Singh, *Introducing Expected Threat (xT) - Modelling team behaviour in possession to gain a deeper understanding of buildup play*, available at `https://karun.in/blog/expected-threat.html` (2019).