

Od podatkov do prvakov: integracija podatkov za podporo odločanja v športu

Jernej Flisar

Twelve Football, Stockholm, Švedska
jernej@twelve.football

Uporaba podatkov je v zadnjih letih postala pomemben del vodenja profesionalnih športnih ekip. Nogomet se je v ta podatkovni razvoj vključil nekoliko pozneje kot nekateri drugi športi, predvsem zaradi same narave igre, ki je dinamična, manj strukturirana in vključuje veliko število igralcev, ki se hkrati gibljejo po velikem prostoru. V prispevku predstavljamo, kako lahko iz surovih nogometnih podatkov pridemo do ključnih informacij, ki jih trenerji, analitiki in drugi člani strokovnega štaba uporabijo v svojih procesih odločanja. Trenerji in analitiki pri tem niso samo zunanji opazovalci, temveč so aktivno vključeni v proces oblikovanja metrik in modelov. Osrednji del prispevka predstavlja praktičen primer, v katerem smo trenerjev princip obrambne kompaktnosti prevedli v podatkovni model in oblikovali metrike, s katerimi je bilo mogoče spremljati, ali ekipa dogovorjene principe dejansko izvaja na igrišču in kje prihaja do odstopanj. Za obdelavo, transformacijo in modeliranje podatkov smo uporabili kombinacijo sistemov DuckDB in MotherDuck ter programskega jezika Python, rezultate pa predstavili prek interaktivnih pregledov v Streamlitu in samodejno pripravljenih poročil v formatu PDF. S prispevkom želimo pokazati, da podatki v nogometu ne nadomeščajo trenerjevega znanja, temveč ga dopolnjujejo tako, da taktične ideje naredijo merljive, preverljive in uporabne v vsakodnevnem delu.

Ključne besede:

nogomet

dogodkovni in sledilni podatki

taktična analiza

DuckDB

podpora odločanju

1 Uvod

Uporaba podatkov je v zadnjih letih postala pomemben del vodenja profesionalnih športnih ekip. Medtem ko so se nekateri športi, predvsem bejzbol in košarka, na podatkovno analitiko naslonili razmeroma zgodaj, se je nogomet v ta podatkovni razvoj vključil nekoliko pozneje, predvsem zaradi same narave igre. Igra je dinamična, manj strukturirana in vključuje veliko število igralcev, ki se hkrati gibljejo po velikem prostoru, zato zbiranje, obdelava in interpretacija podatkov v nogometu predstavljajo poseben izziv. Igra je le redko razdeljena na jasno ločene in ponovljive situacije, ki bi jih bilo mogoče neposredno prešteti in primerjati, posamezni dogodek z žogo pa pove razmeroma malo o tem, kaj se je v istem trenutku dogajalo na preostalem delu igrišča [1].

Dodatna posebnost nogometa je majhno število zadetkov. Ker je zadetkov na tekmo malo, je izid močno odvisen od naključja, kar pomeni, da rezultat sam po sebi ni zanesljiv pokazatelj kakovosti prikazane igre. Prav zato so se v nogometu uveljavile metrike, ki namesto izida merijo proces, ki do izida pripelje, in s tem omogočajo bolj stabilno oceno uspešnosti ekipe skozi daljše obdobje.

V zadnjih letih se je dostopnost podatkov bistveno spremenila. Poleg dogodkovnih podatkov, ki opisujejo akcije z žogo, so postali široko dostopni tudi sledilni podatki, ki zajemajo položaje vseh igralcev in žoge v kratkih časovnih intervalih. Takšni podatki omogočajo analizo kolektivnega taktičnega vedenja ekipe, torej tistega dela igre, ki se dogaja brez žoge in je bil prej dostopen le z videoanalizo in subjektivno presojo [2]. Sistematični pregledi raziskav kažejo, da prav pozicijski podatki odpirajo največ novih možnosti za razumevanje kolektivnega vedenja ekip, hkrati pa opozarjajo, da metodološka raznolikost in pomanjkanje standardizacije otežujeta prenos raziskovalnih ugotovitev v vsakodnevno delo strokovnih štabov [3].

Prav ta prenos je po naših izkušnjah osrednji izziv. Tehnične možnosti za izračun naprednih metrik so danes dostopne tudi manjšim klubom, vendar analitika pogosto ostane ločena od dela strokovnega štaba. Metrike, ki niso povezane z igralnim modelom ekipe in s konkretnimi vprašanji trenerjev, se v praksi ne uporabijo, ne glede na to, kako dovršena je njihova statistična osnova. Dodatno omejitev predstavlja časovni okvir: ker se tekme igrajo vsakih nekaj dni, so roki za pripravo analiz zelo kratki, zahteve pa se hitro spreminjajo glede na nasprotnika, stanje ekipe in ugotovitve iz prejšnjih tekem.

V prispevku najprej podrobneje opišemo glavne sklope nogometnih podatkov, razlike med njimi, njihove ponudnike in osnovne metrike, ki jih iz njih izpeljemo. Nato predstavimo področja, na katerih klubi podatke uporabljajo v vsakdanjem delu, postopek prevajanja taktičnih principov v merljive metrike in podatkovno arhitekturo, na kateri rešitev temelji. Osrednji del prispevka predstavlja študija primera iz sodelovanja s strokovnim štabom kluba angleške Premier lige.

2 Nogometni podatki

Nogometne podatke najpogosteje delimo v tri glavne sklope: dogodkovne, sledilne in fizične. Sklopi se med seboj dopolnjujejo in odgovarjajo na različne vrste vprašanj, zato izbira podatkovnega vira ni tehnična, temveč izhaja iz vprašanj, na katera želimo odgovoriti.

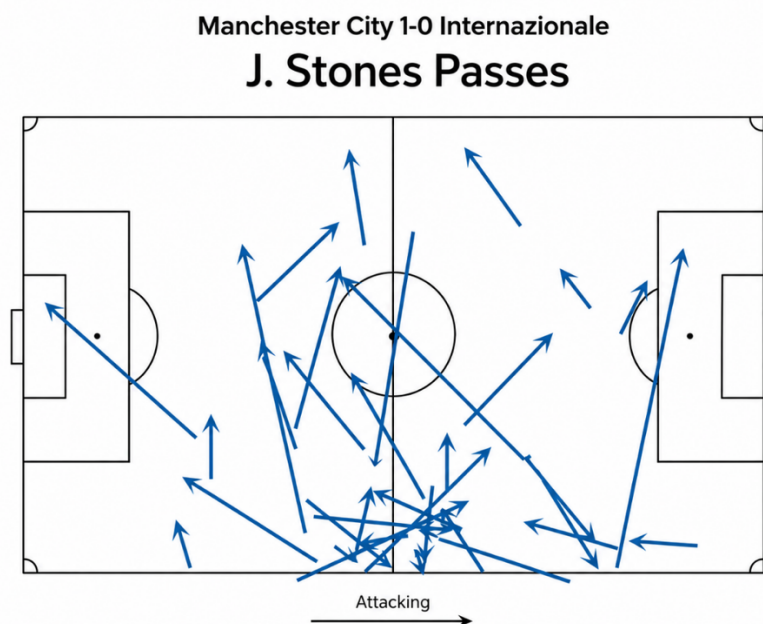
2.1. Dogodkovni podatki

Dogodkovni podatki (angl. event data) opisujejo posamezne akcije akterjev (igralcev) z žogo. Vključujejo podaje, strele, preigravanja, odvzeme žoge, prekrške, prekinitve in zadetke, vsaka akcija pa je umeščena v prostor na igrišču in v čas tekme.

Posamezen zapis običajno vsebuje vrsto dogodka, igralca, ekipo, časovno oznako, koordinate na igrišču in izid akcije, pogosto pa tudi dodatne attribute, na primer telesni del, s katerim je bila akcija izvedena, vrsto podaje ali oznako, ali je bila akcija izvedena pod pritiskom. Na eni tekmi je zabeleženih približno dva do tri tisoč dogodkov.

Zajem je bil dolgo pretežno ročen: operaterji so ob spremljanju posnetka označevali dogodke po vnaprej določenem klasifikacijskem sistemu. V zadnjih letih se ročno označevanje vse bolj dopolnjuje s samodejnim prepoznavanjem, kar znižuje stroške in skrajšuje čas do razpoložljivosti podatkov.

Ključna omejitev dogodkovnih podatkov je, da opisujejo le akterja z žogo. O tem, kje so bili v istem trenutku preostali igralci, ne povedo ničesar. Ker ima igralec žogo v posesti le kratek čas, dogodkovni podatki zajamejo razmeroma majhen del dogajanja. Za analizo obrambne organizacije, ki je po definiciji kolektivna in se večinoma odvija brez žoge, so zato sami po sebi nezadostni.



Slika 1: Dogodkovni podatki – podaje posameznega igralca na eni tekmi. Vsaka puščica prikazuje mesto, smer in dolžino podaje.

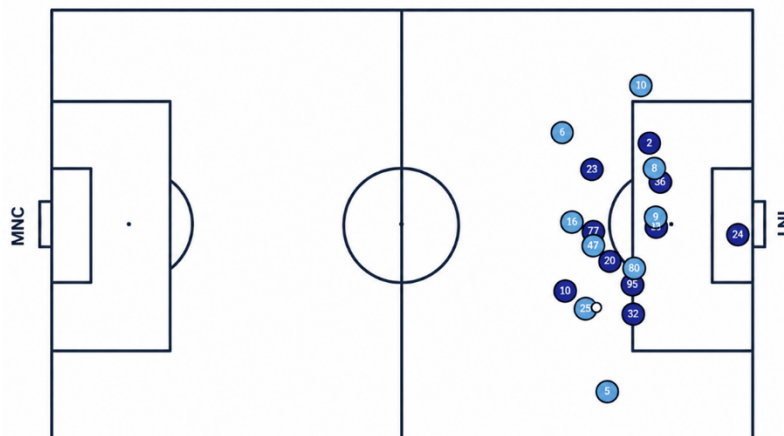
2.2. Sledilni podatki

Sledilni podatki (angl. tracking data) zajemajo položaje, hitrosti in pospeške vseh igralcev ter žoge v kratkih časovnih intervalih, običajno petindvajsetkrat na sekundo. Ena tekma tako obsega več milijonov podatkovnih točk, kar je za nekaj velikostnih redov več od dogodkovnih podatkov in ima neposredne posledice za izbiro podatkovne arhitekture.

Zajem poteka na tri načine. Optični sistemi s kamerami, nameščenimi na stadionu, zagotavljajo najvišjo kakovost in zajamejo vse igralce skozi celotno tekmo. Sistemi, ki sledenje izračunajo iz televizijskega signala (angl. broadcast tracking), so cenejši in dostopni za bistveno več tekmovalj, vendar zajamejo le tiste igralce, ki so v danem trenutku v kadru, zato je treba manjkajoče položaje ocenjevati. Naprave GPS in druge nosljive naprave zagotavljajo natančne podatke o gibanju posameznega igralca, vendar praviloma le za lastno ekipo in predvsem za namene spremljanja obremenitve.

Prav sledilni podatki omogočajo analizo kolektivnega taktičnega vedenja: medsebojnih razdalj med igralci, oblike in višine obrambne linije, kompaktnosti ekipe, gostote pokrivanja prostora ter usklajenosti gibanja. Pregled raziskav o analizi obrambne igre s sledilnimi podatki kaže, da metrike segajo od individualne ravni, kamor sodita obrambni pritisk in razdalja do žoge, do ekipne ravni, kjer se merijo razdalje med linijami, številčna premoč in kompaktnost [4]. Na sledilnih podatkih temeljijo tudi prostorski modeli, ki ocenjujejo, kateri del igrišča posamezna ekipa v danem trenutku obvladuje (angl. pitch control) [5]. Kot je zapisal Johan Cruyff, igralca opredeljuje predvsem to, kaj počne v tistih minutah, ko žoge nima; prav ta del igre postane merljiv šele s sledilnimi podatki.

Manchester City 1-0 Internazionale



Slika 2: Sledilni podatki – položaji vseh igralcev obeh ekip in žoge v enem samem trenutku tekme.

2.3. Fizični podatki

Iz sledilnih podatkov in nosljivih naprav izpeljemo fizične podatke o aktivnosti igralcev: pretečeno razdaljo po hitrostnih conah, najvišjo hitrost, število sprintov, število tekov z visoko intenzivnostjo in število pospeškov. Dopolnjujejo jih meritve srčnega utripa in drugi podatki, ki jih spremlja kondicijski del strokovnega štaba.

Fizične podatke je smiselno vedno obravnavati ločeno za faze s posestjo in za faze brez nje, saj sta profil obremenitve in njegova razlaga v teh dveh fazah povsem različna. Sama pretečena razdalja je brez taktičnega konteksta zavajajoča: večja razdalja ni nujno znak boljše igre, temveč je lahko posledica slabše postavitve, zaradi katere mora igralec pokrivati več prostora.

2.4. Razlike med dogodkovnimi in sledilnimi podatki

Razlika med obema sklopoma ni le v obsegu, temveč predvsem v tem, na katera vprašanja lahko odgovorita. Dogodkovni podatki opisujejo, kaj se je zgodilo z žogo, sledilni podatki pa prikazujejo položaj vseh igralcev v trenutku posameznega dogodka. Zaradi različnih informacij, ki jih zagotavljata, ima vsak sklop določene omejitve. Šele njuna kombinacija omogoča celovitejšo analizo igre. Primerjavo povzemamo v tabeli 1.

Tabela 1: Primerjava dogodkovnih in sledilnih podatkov ter kombinacija obeh.

	Dogodkovni podatki		Sledilni podatki	Dogodkovni in sledilni podatki skupaj
Kaj zajemajo	akcije z žogo (angl. on-ball), ki se zgodijo med tekmo		položaje, hitrosti in pospeške vseh igralcev ter žoge med tekmo	združitev obeh sklopov v enotno predstavitev
Primeri vsebine	zadetki, streli, odvzemi, prekrški	podaje,	gibanje ekipe brez žoge, razdalje med linijami, pokrivanje prostora	akcija z žogo skupaj s postavitvijo vseh igralcev v istem trenutku
Način zajema	označevanje posnetka, bolj samodejno	vse	zajem iz TV signala, kamer na stadionu, naprav GPS	povezava obeh virov po času in identiteti igralcev
Obseg na tekmo	2–3 tisoč dogodkov		več kot 2 milijona podatkovnih točk	obseg sledilnih podatkov, obogaten z oznakami dogodkov
Česa ne pove	kaj se v istem trenutku dogaja s preostalimi igralci		katera akcija z žogo v danem trenutku poteka	

Povezava obeh sklopov je v praksi pogosta, saj šele ta omogoča vprašanja, na katera noben vir sam ne odgovori: koliko nasprotnikov je bilo v bližini podajalca, ali je bil strel izveden pod pritiskom, kateri soigralci so bili v tistem trenutku prosti in katera podajna možnost je ostala neizkoriščena. Nekateri ponudniki takšen kontekst ponujajo neposredno, tako da vsakemu dogodku priložijo položaje igralcev v tistem trenutku, kar predstavlja vmesno obliko med obema sklopoma.

Sklopa se razlikujeta tudi po dostopnosti. Dogodkovni podatki so na voljo za bistveno več tekmovalj, vključno z nižjimi ligami, in so zato običajna osnova za skavtiranje v širokem naboru trgov. Sledilni podatki so dražji in omejeni predvsem na višja tekmovalja, so pa nepogrešljivi za taktično in fizično analizo, kakršno opisujemo v študiji primera.

2.5. Ponudniki podatkov

Trg podatkov je razdeljen med več ponudnikov, ki se razlikujejo po pokritosti tekmovalj, načinu zajema in podrobnosti klasifikacijskega sistema. Med ponudniki dogodkovnih podatkov so najbolj uveljavljeni Stats Perform s podatki pod znamko Opta, StatsBomb in Wyscout, pri čemer se razlikujejo predvsem po številu zabeleženih atributov na dogodek in po pokritosti nižjih tekmovalj.

Pri sledilnih podatkih ločimo optične sisteme, nameščene na stadionih, med katerimi so Second Spectrum, TRACAB in Hawk-Eye, ter sisteme, ki sledenje izračunajo iz televizijskega signala, med katerimi je najbolj razširjen SkillCorner. Prvi zagotavljajo večjo natančnost in popolno pokritost igralcev, drugi pa bistveno širšo pokritost tekmovalj ob nižjih stroških. Fizične podatke s treningov in tekem zagotavljajo proizvajalci nosljivih naprav, med katerimi sta najbolj razširjena Catapult in STATSports.

Za klub izbira ponudnika ni le stroškovna odločitev. Formati in klasifikacijski sistemi se med ponudniki razlikujejo, zato menjava vira pomeni, da zgodovinskih podatkov ni mogoče neposredno primerjati z novimi. Prav to je eden od razlogov, da v podatkovni arhitekturi, opisani v poglavju 5, podatke različnih ponudnikov čim prej pretvorimo v enotno interno shemo.

2.6. Osnovne metrike

Iz opisanih podatkov gradimo metrike, ki surove zapise pretvorijo v oceno vrednosti.

Pričakovani zadetki (angl. expected goals, xG) so najbolj uveljavljena metrika v nogometu in za večino klubov prvi korak v podatkovno analitiko. Model vsakemu strelu pripiše verjetnost, da se konča z zadetkom, pri čemer se opira na okoliščine, ki so bile ob strelu znane: razdaljo in kot do gola, del telesa oziroma način izvedbe, vrsto predhodne akcije, položaj branilcev in vratarja ter podatek, ali je bil strel izveden pod pritiskom. Vrednost xG je torej ocena, kako pogosto bi povprečen igralec iz enake situacije dosegel zadetek: strel z bele točke ima približno 0,76 xG, strel z roba kazenskega prostora pa pogosto manj kot 0,05 xG.

Slika 3 prikazuje, kako verjetnost zadetka pada z oddaljenostjo in manjšim kotom. Območja z enako verjetnostjo tvorijo krivulje okrog gola, pri čemer je razlika med strelom iz osrčja kazenskega prostora in strelom z njegovega roba bistveno večja, kot se zdi na prvi pogled.

Vsota vrednosti xG čez tekmo je precej bolj stabilen pokazatelj kakovosti ustvarjenih priložnosti kot število doseženih zadetkov, saj ni odvisna od naključja pri zaključku. Prav zato xG omogoča ločevanje med kakovostjo igre in trenutno učinkovitostjo: ekipa, ki dosega manj zadetkov, kot znaša njena vrednost xG, praviloma ni igrala slabo, temveč je imela slabši izkoristek, kar se v daljšem obdobju običajno izravna. Enaka logika velja v obrambni fazi, kjer vrednost xG nasprotnika pove, kako nevarne priložnosti je ekipa dopustila.

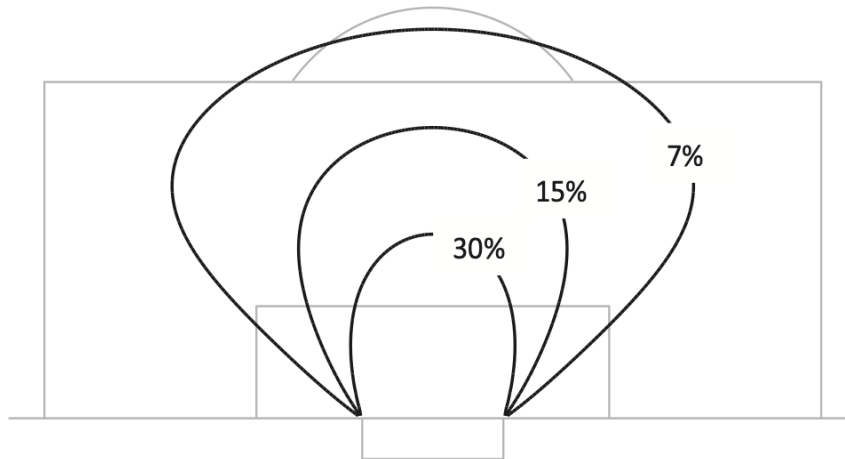
Pričakovane asistenc (angl. expected assists, xA) na enak način ovrednotijo podajo pred strelom in omogočajo, da ustvarjalnost podajalca ločimo od uspešnosti soigralca pri zaključku. Igralec ima lahko visoko vrednost xA brez ene same asistenc, kar pomeni, da redno ustvarja priložnosti, ki jih soigralci ne izkoristijo, ali obratno.

Pričakovana nevarnost (angl. expected threat, xT) razširi vrednotenje strelav in asistenc na napredovanje žoge. Vsaki podaji ali vodenju pripiše spremembo verjetnosti, da bo posest pripeljala do zadetka, in s tem omogoči

merjenje progresivnosti [6]. Metrika pokaže tudi primere, ko ekipa žogo redno prenaša v nevarna območja, a si iz tega ne ustvari priložnosti.

Isto logiko posploši pristop VAEP, ki vsaki akciji na igrišču pripiše spremembo verjetnosti, da bo ekipa v kratkem dosegla oziroma prejela zadetek, in z enotnim merilom ovrednoti tako napadalne kot obrambne akcije [7].

Poleg vrednotenja akcij so v uporabi tudi metrike, ki opisujejo **slog igre**. Metrika PPDA, ki meri število nasprotnikovih podaj na eno obrambno akcijo, na primer pove, kako visoko in kako agresivno ekipa pritiska. Na strani sledilnih podatkov med osnovne metrike sodijo kompaktnost ekipe, višina obrambne linije, razdalje med linijami in nadzor prostora.



Slika 3: Model pričakovanih zadetkov – krivulje enake verjetnosti zadetka glede na oddaljenost in kot strela.

3 Kako klubi uporabljajo podatke

Po naših izkušnjah pri sodelovanju s klubi različnih velikosti in ravni tekmovanja se uporaba podatkov osredotoča na štiri področja, ki se razlikujejo po časovnem okviru in po tem, kdo je naročnik analize.

Analiza lastne ekipe je najbolj dolgoročna in najtesneje povezana z igralnim modelom. Njeno izhodišče ni vprašanje, kaj je v podatkih razvidno, temveč kaj klub sploh želi meriti: kakšen slog igre zasleduje, koliko posesti želi imeti, kako visoko želi pritiskati in kako želi izvajati prehode med posameznimi igralnimi fazami. Šele iz teh odgovorov je mogoče izpeljati smiselne ključne kazalnike uspešnosti. Bistveno je razlikovanje med metrikami, ki merijo slog, in tistimi, ki merijo rezultat: PPDA ne pove, ali ekipa pritiska dobro, temveč le, ali sploh pritiska in kako visoko. Kadar se ti dve vrsti metrik zamešata, se v ocenjevanje ekipe vgradi napačna povratna informacija.

Analiza tekme in nasprotnika je najkrajši in najbolj ponovljiv cikel, saj se ponovi vsakih nekaj dni. Pri analizi nasprotnika je prvo vprašanje, katere njegove tekme sploh pregledati; smiselne so predvsem tiste proti podobnim nasprotnikom in tiste, ki napovedujejo, kako bo nasprotnik igral prav proti naši ekipi. Zanima nas, kako gradi igro in po katerih conah prihaja v napadalno tretjino, kje si ustvarja priložnosti in kje so si priložnosti proti njemu ustvarili drugi.

Tehnično skavtiranje podatke uporablja najprej za zožitev trga. Izhodišče je profil, ki opredeli, katere kakovosti so za določeno vlogo bolj in katere manj pomembne, in šele nato sledi iskanje igralcev, ki temu profilu ustrezajo. Ključno vprašanje ni le, kako dober je igralec, temveč v kakšni ekipi in v kakšnem slogu je igral doslej ter kako bi se njegove kakovosti prenesle v našo ekipo in v naše prvenstvo.

Vsem področjem je skupno, da podatki sami po sebi ne dajo odgovora. Dajo ga šele v kombinaciji z vprašanjem, ki ga postavi strokovni štab, in s kontekstom, ki ga prispeva nogometno znanje.

4 Od igralnega modela do metrik

Izhodišče našega dela ni bil nabor razpoložljivih podatkov, temveč igralni model ekipe. Strokovni štab je imel jasno predstavo o tem, kako naj ekipa deluje, in to predstavo je znal ubesediti v nogometnem jeziku. Naloga podatkovnega dela ni bila predlagati nove koncepte, temveč obstoječe taktične principe prevesti v merljivo obliko, ki jo je mogoče spremljati skozi čas.

Ta prevod poteka v več korakih. Najprej princip skupaj s trenerji razgradimo na merljive elemente: kaj natančno naj bi bilo vidno na igrišču, ko ekipa princip izvaja pravilno, in kaj, ko ga ne. Nato določimo, v katerih fazah igre je princip sploh relevanten, saj je vsaka obrambna metrika smiselna le v ustreznem kontekstu posesti in položaja žoge. Šele nato izberemo podatkovni vir in opredelimo konkreten izračun.

Ključno pri tem je, da trenerji in analitiki niso samo zunanji opazovalci, temveč so aktivno vključeni v proces oblikovanja metrik in modelov. Te morajo biti povezane z igralnim modelom, cilji ekipe in konkretnimi vprašanji strokovnega štaba. Vsako vmesno različico smo predstavili strokovnemu štabu na konkretnih posnetkih tekem in preverili, ali se vrednost metrike ujema s trenersko presojo iste situacije. Kadar se ni, to praviloma ni pomenilo, da se trener moti, temveč da opredelitev metrike še ne zajema nečesa, kar je za presojo bistveno. Takšen iterativen postopek je bistveno povečal verjetnost, da bo metrika pozneje dejansko uporabljena, saj so jo člani štaba razumeli in ji zaupali.

Drugi pomemben dejavnik je hitrost. Ker se tekme igrajo vsakih nekaj dni, mora biti celotna pot od prejema podatkov do pripravljenega pregleda dovolj kratka, da rezultati pridejo do trenerjev, preden se pozornost preusmeri na naslednjega nasprotnika.

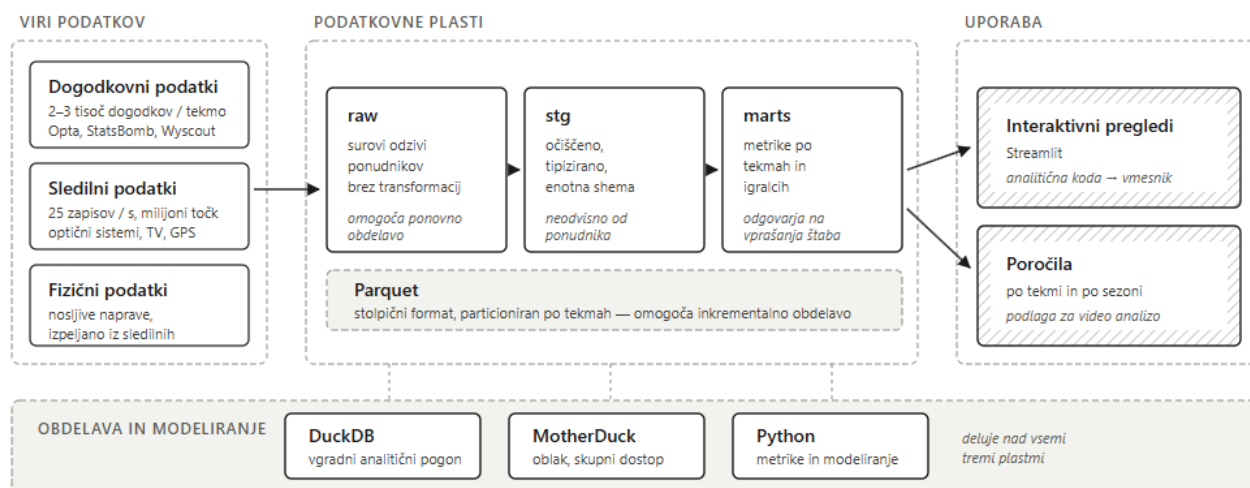
5 Podatkovna arhitektura

Za obdelavo, transformacijo in modeliranje podatkov uporabljamo kombinacijo sistemov DuckDB in MotherDuck ter programskega jezika Python. Izbira je posledica dveh zahtev: obsega sledilnih podatkov, ki na ravni sezone hitro preseže milijardo vrstic, in potrebe po hitri iteraciji, opisani v prejšnjem poglavju.

Sorodne arhitekturne vzorce so na tej konferenci obravnavali že drugi avtorji. Nered in Jeraj sta predstavila podatkovno platformo za zajem biosenzoričnih podatkov v realnem času, ki obravnava podoben razred problemov na področju športa in zdravja [8]. Sestak in Turkanović sta obravnavala nadgradnjo podatkovnih jezer s formatom Delta Lake in razlike med odprtimi formati za shranjevanje v podatkovnem kolišču [9], Heričko s soavtorji pa vpeljavajo tabelaričnih tokov v podatkovno arhitekturo [10]. Naša rešitev je v primerjavi s temi pristopi namerno preprostejša: ker gre za analitično obdelavo omejenega števila virov brez zahtev po sprotni obdelavi v realnem času, se izognemo porazdeljenim okvirjem in ostanemo pri vgrajenem pogonu nad datotekami v objektni shrambi.

DuckDB služi kot učinkovit analitični SQL-pogon za lokalno obdelavo večjih količin podatkov. Gre za vgrajeni podatkovni sistem, ki poizvedbe izvaja v procesu same aplikacije in je zasnovan za stolpčno analitično obdelavo [11], zaradi česar omogoča neposredno poizvedovanje po datotekah v formatu Parquet brez predhodnega nalaganja v ločen strežnik. MotherDuck isti pogon razširja v oblako okolje za shranjevanje, deljenje in izvajanje analiz, kar pomeni, da analitiki in drugi člani štaba dostopajo do istih pripravljenih tabel brez lokalne namestitve [12]. Python uporabljamo za izračun naprednejših metrik, modeliranje ter pripravo vizualizacij in poročil. Takšna arhitektura je omogočila hitro iteracijo od surovih prostorsko-časovnih podatkov do uporabnih vpogledov brez potrebe po kompleksni klasični podatkovni infrastrukturi.

Podatke organiziramo v tri plasti. V plasti *raw* hranimo surove odzive ponudnikov v izvorni obliki, brez transformacij, kar omogoča ponovno izvedbo celotne obdelave, kadar se opredelitev metrike spremeni. V plasti *stg* so podatki očiščeni, tipizirani in normalizirani v enotno shemo, neodvisno od ponudnika; ta korak je ključen prav zaradi razlik med ponudniki, opisanih v poglavju 2.5. V plasti *marts* so pripravljene tabele, ki neposredno odgovarjajo na vprašanja strokovnega štaba, na primer vrednosti metrik po tekmah in igralcih. Vsaka plast je shranjena v stolpcnem formatu Parquet [13] in razdeljena na naravne particije, običajno po tekmah, kar omogoča inkrementalno obdelavo: ob novi tekmi ni treba ponovno obdelati celotne sezone. Celotno arhitekturo prikazuje slika 4.



Slika 4: Arhitektura podatkovne rešitve od virov podatkov prek podatkovnih plasti do uporabe.

Rezultate predstavljamo prek interaktivnih pregledov v Streamlitu in samodejno pripravljenih poročil v formatu PDF. Streamlit omogoča, da se analitična koda v Pythonu neposredno pretvori v uporabniški vmesnik brez ločenega razvoja spletne aplikacije [14]. Poročila pripravljamo v dveh oblikah: poročilo o posamezni tekmi, namenjeno pripravi na videoanalizo v dneh po tekmi, in sezonsko poročilo, ki prikazuje razvoj metrik v daljšem obdobju. To je bistveno skrajšalo pot od nove ideje do pregleda, ki si ga lahko ogleda strokovni štab, in omogočilo, da so vmesne različice metrik nastajale v tesnem sodelovanju s trenerji.

Med razvojem se je pokazalo, da je prav ločitev plasti *raw* in *stg* tista, ki omogoča pogoste spremembe opredelitev metrik, saj lahko izračune ponovimo nad že zajetimi podatki, ne da bi jih morali znova pridobivati od ponudnika.

6 Študija primera: obrambna kompaktnost

Osrednji del prispevka predstavlja praktičen primer uporabe podatkov pri razvoju metrike, ki povezuje igralni model ekipe z njenimi taktičnimi principi. Sodelovali smo s strokovnim štabom kluba angleške Premier lige v okviru nekaj tekmovalnih sezon; primer, ki ga opisujemo v nadaljevanju, izhaja iz ene izmed njih.

6.1 Taktični princip

Strokovni štab je imel jasno predstavo o tem, kako želi, da ekipa deluje v obrambi, predvsem kako povezano mora biti gibanje obrambnih igralcev glede na položaj žoge in nasprotnika. Princip je bil ubeseden v obliki preprostega pravila: branilci naj v fazi brez žoge od žoge nikoli ne bodo oddaljeni več kot tri navpične pasove igrišča (angl. lanes). Pravilo je bilo igralcem razumljivo in ga je bilo preprosto posredovati, hkrati pa je bilo dovolj natančno, da ga je bilo mogoče prevesti v merljivo obliko. To nogometno idejo smo prevedli v podatkovni model in oblikovali metrike, s katerimi je bilo mogoče spremljati, ali ekipa dogovorjene principe dejansko izvaja na igrišču, kje prihaja do odstopanj in ali se skozi čas približuje zelenemu načinu igre.

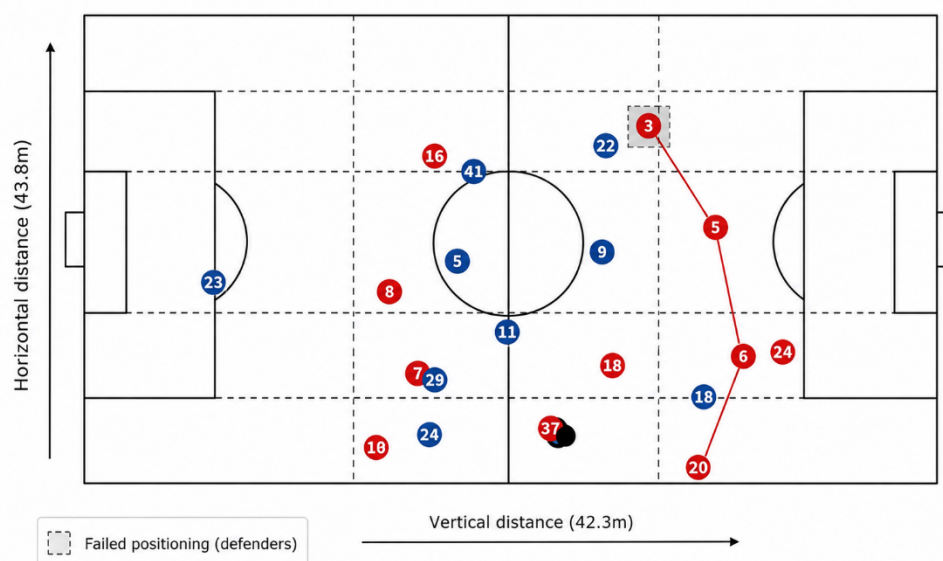
Prevelika oddaljenost obrambne skupine od žoge pomeni, da se odpre prostor, ki ga lahko nasprotnik izkoristi za prodor ali za podajo za hrbet obrambne linije, hkrati pa oteži medsebojno pokrivanje branilcev.

Trenerjeva zahteva se dobro ujema z ugotovitvami raziskav. Analiza sledilnih podatkov nemške Bundeslige je pokazala, da je bila skupina petih branilcev, ki so bili najbližje žogi, ob uspešnih obrambnih akcijah bolj kompaktna kot ob neuspešnih, medtem ko kompaktnost celotne ekipe za odvzem žoge ni bila pomembna [15]. Merjenje kompaktnosti glede na položaj žoge in ne na celotno ekipo je torej utemeljeno tudi z vidika raziskav, kar je bil dodaten argument pri izbiri definicije metrike.

6.2. Definicija metrike

Princip smo operacionalizirali kot **horizontalno kompaktnost obrambne skupine**, merjeno v metrih in izračunano izključno v fazah, ko ekipa ni imela posesti žoge. Metrika meri, koliko je obrambna skupina raztegnjena po širini igrišča, torej med najbolj oddaljenima bočnima branilcema; nižja vrednost pomeni bolj kompaktno postavitvev. Prav ta razdalja neposredno ustreza trenerjevemu pravilu o treh navpičnih pasovih, saj se pasovi merijo po širini igrišča. Izračunali smo jo iz sledilnih podatkov, ki zagotavljajo položaje vseh igralcev skozi celotno tekmo [16].

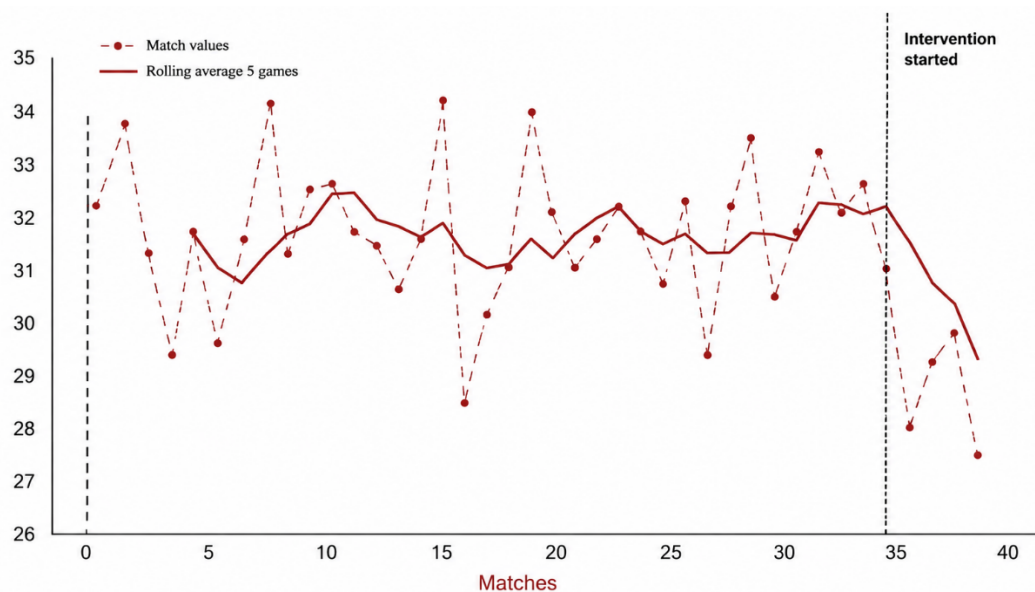
Poleg agregirane vrednosti smo razvili tudi diagnostični pogled, ki posamezne trenutke označi kot primere neustrezne postavitve (angl. failed positioning). Za vsak tak trenutek prikažemo postavitve obeh ekip na igrišču skupaj s horizontalno in navpično razdaljo obrambne skupine, kar omogoča, da se agregirana številka takoj poveže s konkretno situacijo na posnetku. Prav ta povezava med vrednostjo in posnetkom se je v pogovorih s štabom izkazala za bistveno, saj trenerji ne delajo z metrikami, temveč z nogometnimi situacijami.



Slika 5: Prikaz trenutka z neustrezno postavitvijo obrambne skupine ter horizontalno in navpično razdaljo med igralci.

6.3. Spremljanje skozi sezono

Metriko smo po vsaki tekmi izračunali in jo prikazali v kontekstu celotne sezone. Poleg vrednosti posamezne tekme smo prikazovali drseče povprečje petih tekem, saj posamična tekma niha glede na nasprotnika in potek srečanja, trajnejši vzorec pa je razviden šele iz daljšega obdobja. Prav ločevanje med posamičnim odstopanjem in trajnejšim vzorcem je bilo za strokovni štab najbolj uporabno.



Slika 6: Sezonski trend horizontalne kompaktnosti v fazi brez žoge z drsečim povprečjem petih tekem in označeno točko intervencije.

6.4. Ugotovitev in intervencija

Spremljanje postavitev posameznih igralcev je pokazalo, da je bil eden od branilcev v fazah brez žoge sistematično postavljen predaleč od žoge glede na preostanek obrambne skupine. Odstopanje ni bilo posledica posamične napake, temveč se je ponavljalo skozi več tekem, kar je bilo v podatkih razvidno kot vztrajno povečana horizontalna kompaktnost v določenih fazah igre.

Strokovni štab je ugotovitev uporabil neposredno v delu z igralcem. Pri videoanalizi so mu prikazali konkretne situacije iz diagnostičnega pogleda in mu pojasnili, kakšna postavitev se od njega pričakuje glede na položaj žoge. Po tej intervenciji se je vrednost metrike opazno znižala (slika 6). Ekipa je v fazi brez žoge torej postala bolj kompaktna in bližje želenemu izvajanju dogovorjenega principa.

Ta primer po našem mnenju najbolje ponazarja vlogo podatkov v opisanem procesu. Metrika sama po sebi ni odkrila ničesar, česar trenerji ne bi mogli opaziti tudi z videoanalizo. Njena vrednost je bila v tem, da je vzorec pokazala sistematično in količinsko, na več tekmah hkrati, in s tem omogočila, da je bil pogovor z igralcem podprt s konkretnim dokazom namesto s splošnim vtisom. Prav tako je omogočila, da je bil učinek intervencije po njeni izvedbi tudi izmerjen. K temu je treba dodati še obseg: metriko izračunamo za vse tekme sezone hkrati, medtem ko bi ročno pregledovanje posnetkov za enak nabor zahtevalo več ur dela, zato takšna odstopanja prej opazimo.

7 Zaključek

V prispevku smo predstavili celotno pot od surovih nogometnih podatkov do konkretne odločitve strokovnega štaba. Opisali smo glavne sklope nogometnih podatkov, razlike med dogodkovnimi in sledilnimi podatki, ponudnike in osnovne metrike, področja, na katerih klubi podatke uporabljajo, postopek prevajanja taktičnih principov v merljive metrike, podatkovno arhitekturo, zgrajeno na kombinaciji DuckDB, MotherDuck in Pythona, ter študijo primera, v kateri smo trenerjev princip obrambne kompaktnosti operacionalizirali kot horizontalno kompaktnost obrambne skupine v fazah brez žoge.

Izbrana arhitektura se je izkazala za primerno. Ločitev podatkov v plasti *raw*, *stg* in *marts* je omogočila, da smo opredelitev metrike med razvojem večkrat spremenili in izračune ponovili nad že zajetimi podatki, particioniranje po tekmah pa je omogočilo inkrementalno obdelavo v časovnem okviru, ki ga narekuje tekmovalni ritem. Za obseg

podatkov, s katerim smo delali, klasična podatkovna infrastruktura ne bi prinesla bistvenih prednosti, prinesla pa bi znatno večjo kompleksnost vzpostavitve in vzdrževanja.

S prispevkom želimo pokazati, da podatki v nogometu ne nadomeščajo trenerjevega znanja, temveč ga dopolnjujejo. Njihova vrednost je v tem, da taktične ideje naredijo merljive, preverljive in uporabne v vsakodnevnem delu. V predstavljenem primeru je takšen pristop strokovnemu štabu omogočil spremljanje izvajanja obrambnih principov, prepoznavanje odstopanj in pripravo bolj ciljno usmerjenih sprememb, utemeljenih na podatkih. Prednost podatkov v primerjavi zgolj z videoanalizo ni v tem, da bi podatki pokazali kaj, česar trener ne bi videl, temveč v obsegu in hitrosti: isto vprašanje lahko preverimo na vseh tekmah sezone hkrati v realnem času.

Literatura

- [1] R. Rein in D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," SpringerPlus, let. 5, št. 1410, 2016, doi: <https://doi.org/10.1186/s40064-016-3108-2>.
- [2] F. R. Goes idr., "Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review," European Journal of Sport Science, let. 21, št. 4, str. 481–496, 2021, doi: <https://doi.org/10.1080/17461391.2020.1747552>.
- [3] B. Low idr., "A Systematic Review of Collective Tactical Behaviours in Football Using Positional Data," Sports Medicine, let. 50, str. 343–385, 2020, doi: <https://doi.org/10.1007/s40279-019-01194-7>.
- [4] L. Forcher, S. Altmann, L. Forcher, D. Jekauc in M. Kempe, "The use of player tracking data to analyze defensive play in professional soccer – A scoping review," International Journal of Sports Science and Coaching, 2022, doi: <https://doi.org/10.1177/17479541221075734>.
- [5] J. Fernández in L. Bornn, "Wide Open Spaces: A statistical technique for measuring space creation in professional soccer," v MIT Sloan Sports Analytics Conference, 2018.
- [6] K. Singh, "Introducing Expected Threat (xT)," 2019. [Na spletu]. Dostopno na: <https://karun.in/blog/expected-threat.html>, obiskano [julij 2026].
- [7] T. Decroos, L. Bransen, J. Van Haaren in J. Davis, "Actions Speak Louder than Goals: Valuing Player Actions in Soccer," v Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, str. 1851–1861, doi: <https://doi.org/10.1145/3292500.3330758>.
- [8] U. Nered in J. Jeraj, "Podatki v utripu srca: podatkovna platforma za zajem biosenzoričnih podatkov v realnem času," v Sodobne informacijske tehnologije in storitve OTS 2023: Zbornik 26. konference, Maribor, 2023.
- [9] M. Sestak in M. Turkanović, "Nadgradnja podatkovnih jezer s pomočjo formata za shranjevanje Delta Lake," v Sodobne informacijske tehnologije in storitve OTS 2023: Zbornik 26. konference, Maribor, 2023, doi: <https://doi.org/10.18690/um.feri.8.2023.21>.
- [10] T. Heričko, S. Brdnik in M. Turkanović, "Vpeljava tabelaričnih tokov v podatkovno arhitekturo," v Sodobne informacijske tehnologije in storitve OTS 2025: Zbornik 28. konference, Maribor, 2025, str. 13–24.
- [11] M. Raasveldt in H. Mühleisen, "DuckDB: an Embeddable Analytical Database," v Proceedings of the 2019 International Conference on Management of Data (SIGMOD), 2019, str. 1981–1984, doi: <https://doi.org/10.1145/3299869.3320212>.
- [12] motherduck.com, MotherDuck Documentation, obiskano 15. 2. 2026.
- [13] parquet.apache.org, Apache Parquet Documentation, obiskano 13. 2. 2026.
- [14] streamlit.io, Streamlit Documentation, obiskano 15. 2. 2026.
- [15] L. Forcher, L. Forcher, S. Altmann, D. Jekauc in M. Kempe, "Is a compact organization important for defensive success in elite soccer? – Analysis based on player tracking data," International Journal of Sports Science and Coaching, let. 19, št. 2, str. 757–768, 2024, doi: <https://doi.org/10.1177/17479541231172695>.
- [16] statsperform.com, Opta Tracking Data, obiskano 10. 2. 2026.