Προβλέψεις για το ΠΑΓΚΟΣΜΙΟ ΚΥΠΕΛΛΟ ΠΟΔΟΣΦΑΙΡΟΥ 2022 βασισμένες σε Στατιστικά Μοντέλα Αναλυτικής Ποδοσφαίρου

L. Egidi, Β. Παλάσκας, Ι. Ντζούφρας & Δ. Καρλής

Ερευνητική ομάδα AUEB Sports Analytics,

Οικονομικό Πανεπιστήμιο Αθηνών, Πανεπιστήμιο της Τεργέστης & Fantasy Sports Interactive

Συγγραφέας του παρόντος άρθρου είναι ο Ιωάννης Ντζούφρας. Το άρθρο είναι βασισμένο στην ανάλυση των Leonardo Egidi (Πανεπιστήμιο της Τεργέστης) και του Βασίλειου Παλάσκα (Fantasy Sports Interactive) με τις συμβουλευτικές συνδρομές των Ι. Ντζούφρα και Δ. Καρλή. Και οι τέσσερεις συγγραφείς είναι ενεργά μέλη της ερευνητικής ομάδας του Οικονομικού Πανεπιστημίου Αθηνών: AUEB Sports Analytics Group.

***

Ήρθε η ώρα του Παγκοσμίου κυπέλου 2022 που φέτος διεξάγεται στο Κατάρ. Ένα περίεργο κύπελο καθώς είναι στη μέση του χειμώνα σε αντίθεση με τις προηγούμενες διοργανώσεις που διεξάγονταν κατά τους καλοκαιρινούς μήνες.

Η χρήση στατιστικών τεχνικών για την πρόβλεψη αγώνων ποδοσφαίρου πρώτο-εμφανίστηκε στην επιστημονική βιβλιογραφία το 1968 με την πρωτοπόρα επιστημονική δημοσίευση των Reep & Benjamin. Οι επόμενες πραγματικές καινοτομίες εμφανίζονται στη δεκαετία του 80 (με την εργασία του Michael Maher) και τη δεκαετία του 90 (με την εργασία του Lee το 1997). Οι πρώτες όμως σημαντικές δημοσιεύσεις στο χώρο, εισάγοντας μοντέλα στα οποία βασίζονται και μοντέλα που χρησιμοποιούμε ακόμα και σήμερα, ήταν οι εργασίες των Dixon & Coles το 1997 και το διμεταβλητό μοντέλο Poisson των Καρλή και Ντζούφρα το 2003 (δύο από τους συγγραφείς της συγκεκριμένης ανάλυσης). Τα δύο αυτά μοντέλα έθεσαν τη βάση των συγχρόνων μοντέλων πρόβλεψης των αποτελεσμάτων αγώνων ποδοσφαίρου.

Σε αυτή την ανάλυση χρησιμοποιούμε ακριβώς το μοντέλο των Καρλή και Ντζούφρα μέσω του πακέτου “footbayes” στη στατιστική γλώσσα προγραμματισμού R που έχουν αναπτύξει οι 2 πρώτοι συγγραφείς αυτού του άρθρου και της ανάλυσης. Το μοντέλο επίσης συμπεριλαμβάνει την εκτίμηση παραμέτρων που εκτιμούν την απόδοση κάθε ομάδας που αλλάζουν στον χρόνο. Για την εκμάθηση του μοντέλου χρησιμοποιήθηκαν περισσότερα από 3000 διεθνείς αγώνες της περιόδου 2018-2022. Κύρια επεξηγηματική μεταβλητή είναι η διαφορά μεταξύ των δύο ομάδων στο δείκτη Coca-Cola/FIFA ranking. Το μοντέλο, που προτάθηκε για πρώτη φορά από τους Καρλή &

Ντζούφρα το 2003, επεκτείνει το συνηθισμένο διμεταβλητό μοντέλο Poisson. Λεπτομέρειες για το μοντέλο στατιστικής και μηχανικής μάθησης που χρησιμοποιήθηκε θα βρείτε στο τέλος αυτού του άρθρου.

Απολογισμός 1ης αγωνιστικής

Δυστυχώς, με το παγκόσμιο κύπελο ποδοσφαίρου φέτον να είναι στη μέση του χειμώνα και στο απόγειο των οικονομικών και οικογενειακών υποχρεώσεων μας, ήταν αδύνατο να βρω χρόνο να γράψω αυτό το άρθρο νωρίτερα – και φυσικά ως αποτέλεσμα να μην έχω δει πάρα ελάχιστα λεπτά από το μουντιάλ (κατάφερα όμως και είδα ζωντανά το πέναλτι του Bale και ένα εξαιρετικό γκολ της Βραζιλίας).

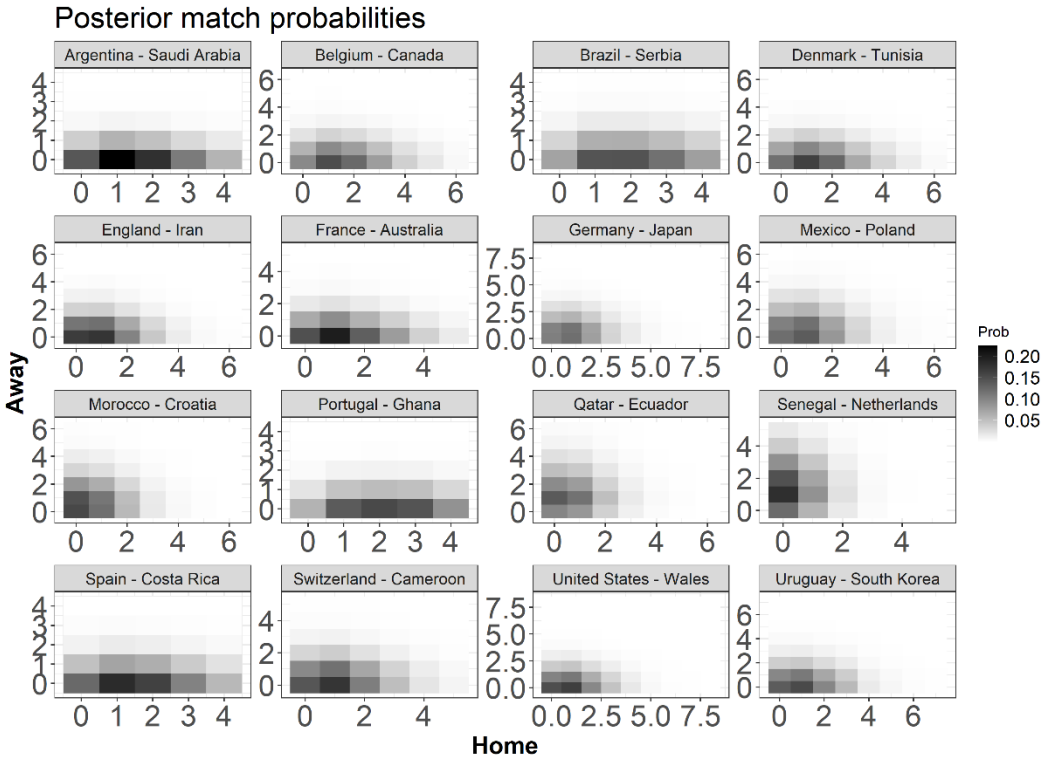
Οπότε εδώ θα γράψω έναν απολογισμό του πόσο καλά πήγε το μοντέλο στο 1ο γύρο. Οι προβλέψεις είχαν αναρτηθεί στην ιστοσελίδα του συνεργάτη μας Leonardo Egidi πριν τους αγώνες φυσικά και είναι ακόμα διαθέσιμες εδώ.

Στον Πίνακα 1 θα βρείτε τις προβλέψεις των πρώτων 16 αγώνων (1η αγωνιστική). Όπως βλέπετε το μοντέλο κατάφερε να προβλέψει σωστά αρκετούς αγώνες (56%) ενώ σε 3 ακόμα αγώνες έδειχνε το τελικό αποτέλεσμα ήταν αρκετά πιθανό. Πιο συγκεκριμένα στον αγώνα Μεξικό-Πολωνία έχουμε αξιοσημείωτη πιθανότητα και στις δύο ομάδες (42% για το Μεξικό έναντι 30% για την Πολωνία) υποδεικνύοντας ότι οι δύο ομάδες είναι κοντά σε δυναμικότητα. Παρόμοια είναι η εικόνα στο Μαρόκο (29%) με την Κροατία (41%) με 30% την πιθανότητα ισοπαλίας. Αξιοσημείωτο είναι ότι το μοντέλο έδινε σημαντική πιθανότητα νίκης στην Ιαπωνία (31%) έναντι της Γερμανίας (42%) στο οποίο κέρδισε η πρώτη. Όσον αφορά την μεγάλη έκπληξη της Σαουδικής Αραβίας, το μοντέλο έδινε μόνο 6% πιθανότητα νίκης όπως και κάθε λογικό μοντέλο θα έδινε. Να σημειώσουμε εδώ ότι ένα λογικό μοντέλο στατιστικής και μηχανικής μάθησης σε καμία περίπτωση δε θα μπορέσει να πιάσει εκπλήξεις σαν και αυτές και μπορούν να συμβούν λόγω απλής τυχαιότητα ή καταστάσεων που δεν λαμβάνονται υπόψη από το μοντέλο.

Πίνακας 1: Πίνακας με τις πιθανότητες έκβασης των αγώνων για την 1η αγωνιστική του Παγκοσμίου Κυπέλου 2022

Αγώνας	Αντίπαλες ομάδες		Όμιλος	Νίκη 1ης ομάδας	Ισοπαλία	Νίκη 2ης ομάδας	Τελικό Αποτέλεσμα
1	Qatar	Ecuador	A	0.229	0.259	0.512	0-2
2	England	Iran	B	0.466	0.311	0.223	6-2
3	Senegal	Netherlands	A	0.115	0.236	0.649	0-2
4	United States	Wales	B	0.474	0.292	0.234	1-1
5	Argentina	Saudi Arabia	Γ	0.718	0.224	0.059	1-2
6	Denmark	Tunisia	Δ	0.592	0.249	0.159	0-0
7	Mexico	Poland	Γ	0.417	0.282	0.301	0-0
8	France	Australia	Δ	0.613	0.261	0.125	4-1
9	Morocco	Croatia	ΣΤ	0.290	0.303	0.407	0-0
10	Germany	Japan	E	0.419	0.269	0.312	1-2
11	Spain	Costa Rica	E	0.686	0.225	0.088	7-0
12	Belgium	Canada	ΣΤ	0.624	0.227	0.148	1-0
13	Switzerland	Cameroon	Z	0.509	0.289	0.202	1-0
14	Uruguay	South Korea	H	0.489	0.280	0.230	0-0
15	Portugal	Ghana	H	0.811	0.144	0.045	3-2
16	Brazil	Serbia	Z	0.748	0.177	0.074	2-0

Το Διάγραμμα 1 δίνει με πιο πολύ λεπτομέρεια τις πιθανότητες για το κάθε σκορ για καθένα από τους πρώτους 16 αγώνες.



Διάγραμμα 1: Διάγραμμα Πιθανοτήτων πιθανών σκορ για τους Αγώνες της πρώτης αγωνιστικής του Παγκοσμίου Κυπέλου 2022.

Οι Προβλέψεις του Μοντέλου για την 2η Αγωνιστική

Οι προβλέψεις για τους αγώνες της 2ης αγωνιστικής δίνονται στον Πίνακα 2. Για τον υπολογισμό τους έχουν ληφθεί υπόψη και τα αποτελέσματα της 1ης αγωνιστικής.

Από τον πίνακα αυτό ξεχωρίζουμε του αγώνες

- Ουαλία – Ιράν
- Τυνησία – Αυστραλία
- Πολωνία – Σαουδική Αραβία

ως τους πιο αμφίρροπους αγώνες.

Ως φαβορί ξεχωρίζουν

1. Βραζιλία με πιθανότητα νίκης 70% έναντι της Ελβετίας
2. Ιαπωνία με πιθανότητα νίκης 65% έναντι της Κοστα Ρίκα
3. Ολλανδία με πιθανότητα νίκης 60% έναντι της Εκουαδόρ
4. Αργεντινή (παρόλο που έχασε τον 1ο αγώνα) με πιθανότητα νίκης 60% έναντι του Μεξικό
5. Νότια Κορέα με πιθανότητα νίκης 59% έναντι της Γκάνα
6. Βέλγιο με πιθανότητα νίκης 59% έναντι του Μαρόκο
7. Ισπανία με πιθανότητα νίκης 55% έναντι της Γερμανίας (και αν συμβεί αυτό η Γερμανία μενεί εκτός της διοργάνωσης).

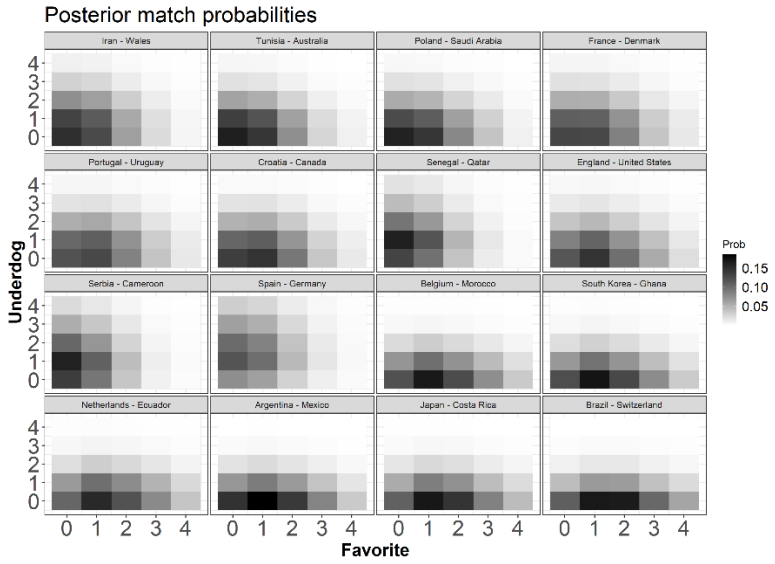
Τέλος έχουμε άλλους έξι αγώνες που είναι σχετικά αμφίρροποι αλλά με ελαφρό προβάδισμα της μίας από τις δύο ομάδες. Σε αυτούς τους αγώνες θεωρούμε ότι οι ομάδες είναι σχετικά κοντά και μπορεί να έρθουν και ισοπαλία λόγω τακτικής και στρατηγικής. Συγκεκριμένα έχουμε

1. Σερβία (50%) να επικρατεί της Καμερούν (22%)
2. Αγγλία (49%) να επικρατεί των ΗΠΑ (23,5%)
3. Σενεγάλη (47%) να επικρατεί του Κατάρ (24%)
4. Κροατία (43%) να επικρατεί του Καναδά (28%)
5. Πορτογαλία (42,5%) να επικρατεί της Ουρουγουάης (29%)
6. Γαλλία (42%) να επικρατεί της Δανίας (29%)

Πίνακας 2: Πιθανότητες Αποτελεσμάτων για τη 2η Αγωνιστική με βάση το Μοντέλο Μπευζιανής Στατιστικής Μηχανικής Μάθησης της Ερευνητικής ομάδας AUEB Sports Analytics

Αγώνας	Αντίπαλες ομάδες		Όμιλος	Νίκη 1ης ομάδας	Ισοπαλία	Νίκη 2ης ομάδας
1	Wales	Iran	Β	0.327	0.305	0.369
2	Qatar	Senegal	Α	0.239	0.292	0.469
3	Netherlands	Ecuador	Α	0.602	0.246	0.152
4	England	United States	Β	0.493	0.272	0.235
5	Tunisia	Australia	Δ	0.368	0.318	0.313
6	Poland	Saudi Arabia	Γ	0.396	0.310	0.294
7	France	Denmark	Δ	0.419	0.288	0.293
8	Argentina	Mexico	Γ	0.605	0.265	0.130
9	Japan	Costa Rica	Ε	0.646	0.235	0.120
10	Germany	Spain	Ε	0.207	0.239	0.554
11	Belgium	Morocco	ΣΤ	0.590	0.258	0.153
12	Croatia	Canada	ΣΤ	0.429	0.294	0.277
13	Cameroon	Serbia	Ζ	0.215	0.285	0.500
14	Brazil	Switzerland	Ζ	0.703	0.212	0.086
15	Portugal	Uruguay	Η	0.425	0.284	0.292
16	South Korea	Ghana	Η	0.594	0.257	0.149

Στο Διάγραμμα 2 μπορείτε να δείτε τις πιθανότητες για το κάθε σκορ για καθένα από τους 16 αγώνες της 2ης αγωνιστικής.



Διάγραμμα 2: Διάγραμμα Πιθανοτήτων πιθανών σκορ για τους Αγώνες της 2ης αγωνιστικής του Παγκοσμίου Κυπέλου 2022.

Βιβλιογραφία για διαβαστερούς φιλάθλους

- Dixon, M.J. and Coles, S.G. (1997), Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**, 265-280.
- Karlis, D. and Ntzoufras, I. (2003), Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381-393.
- Lee A.J. (1997). Modeling Scores in the Premier League: Is Manchester United Really the Best? *Chance*, **10**, 15-19.
- Maher, M.J. (1982), Modelling association football scores. *Statistica Neerlandica*, **36**, 109-118.
- Reep, C., & Benjamin, B. (1968). Skill and Chance in Association Football. *Journal of the Royal Statistical Society. Series A (General)*, **131**, 581-585.

Οι Μαγικές Εξισώσεις του στατιστικού μοντέλου

$$(X_i, Y_i) \sim \begin{cases} (1-p)\mathrm{BP}(x_i, y_i|\lambda_1, \lambda_2, \lambda_3) & \text{if } x \neq y \\ (1-p)\mathrm{BP}(x_i, y_i|\lambda_1, \lambda_2, \lambda_3) + pD(x, \eta) & \text{if } x = y, \end{cases} \quad (1)$$

$$\log(\lambda_{1i}) = \mathrm{att}_{h_i,t} + \mathrm{def}_{a_i,t} + \frac{\gamma}{2}(\mathrm{ranking}_{h_i} - \mathrm{ranking}_{a_i}) \quad (2)$$

$$\log(\lambda_{2i}) = \mathrm{att}_{a_i,t} + \mathrm{def}_{h_i,t} - \frac{\gamma}{2}(\mathrm{ranking}_{h_i} - \mathrm{ranking}_{a_i}), \quad i = 1, \ldots, n \text{ (matches)}, \quad (3)$$

$$\log(\lambda_{3i}) = \rho, \quad (4)$$

$$\mathrm{att}_{k,t} \sim \mathcal{N}(\mathrm{att}_{k,t-1}, \sigma^2), \quad (5)$$

$$\mathrm{def}_{k,t} \sim \mathcal{N}(\mathrm{def}_{k,t-1}, \sigma^2), \quad (6)$$

$$\rho,\ \gamma \sim \mathcal{N}(0, 1) \quad (7)$$

$$p \sim \mathrm{Uniform}(0, 1) \quad (8)$$

$$\sum_{k=1}^{n_t} \mathrm{att}_{k,} = 0,\ \sum_{k=1}^{n_t} \mathrm{def}_{k,} = 0, \quad k = 1, \ldots n_t \text{ (teams)},\ t = 1, \ldots, T \text{ (times)}. \quad (9)$$

- *i* είναι ο δείκτης του αγώνα
- X_i και Y_i είναι ο αριθμός των γκολ μεταξύ της 1ης και της 2ης ομάδας στον αγώνα i
- h_i και a_i είναι η 1η και 2η ομάδα αντίστοιχα (ή η εντός και εκτός έδρα ομάδα – όπου ισχύει) για τον *i* αγώνα.

- $att_{k,t}$ και $def_{k,t}$ οι παράμετροι που εκτιμούν της επιθετική και αμυντική δυναμικότητα/ ικανότητα της ομάδας k την χρονική στιγμή t (δυναμικές παράμετροι που αλλάζουν στο χρόνο)
- $ranking_k$ δείκτης Coca-Cola FIFA ranking την 6η Οκτωβρίου 2022 για την ομάδα k.

Λίγα λόγια για τους Συγγραφείς

Ο ***Leonardo Egidi*** είναι επίκουρος καθηγητής Στατιστικής στο Πανεπιστήμιο της Τεργέστης στην Ιταλία και μέλος της ερευνητικής ομάδας του Οικονομικού Πανεπιστημίου Αθηνών AUEB Sports Analytics Group. Έχει διδακτορικό στην μοντελοποιηση και αναλυτική ποδοσφαίρου και έντονη ερευνητική δραστηριότητα στη Μπευζιανή Στατιστική μεθοδολογία.

Ο ***Βασίλης Παλάσκας*** είναι Στατιστικός Αναλυτής και Επιστήμονας Δεδομένων στην Fantasy Sports Interactive (FSI). Είναι ενεργό μέλος της ερευνητικής ομάδας AUEB Sports Analytics από το 2019 όπου τελείωσε το M.Sc. in Statistics του Οικονομικού Πανεπιστημίου Αθηνών.

Ο ***Ιωάννης Ντζουφρας*** είναι καθηγητής Στατιστικής και πρόεδρος στο Τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών. Είναι ιδρυτικό μέλος της ερευνητικής ομάδας AUEB Sports Analytics Group μαζί με τον Δημήτρη Καρλή. Έχει αναγνωρισμένη επιστημονική δραστηριότητα σε τομείς όπως η Μπευζιανή στατιστική μεθοδολογία, υπολογιστική στατιστική, Βιοστατιστική, ψυχομετρία και αναλυτική των σπορ.

Ο ***Δημήτρης Καρλής*** είναι καθηγητής Στατιστικής και αναπληρωτής πρόεδρος στο Τμήμα Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών. Είναι ιδρυτικό μέλος της ερευνητικής ομάδας AUEB Sports Analytics Group μαζί με τον Ιωάννη Ντζούφρα. Έχει αναγνωρισμένη επιστημονική δραστηριότητα σε τομείς όπως η στατιστική μεθοδολογία, υπολογιστική στατιστική, Βιοστατιστική, και αναλυτική των σπορ.

Ενεργές Συνεργασίες των Συγγραφέων

Οι τρεις συγγραφείς (L. Egidi, I. Ντζούφρας και Δ. Καρλής) του άρθρου αυτή τη στιγμή συνεργάζονται για τη συγγραφή ενός βιβλίου σε ***Football Analytics*** σε διεθνή επιστημονικό οίκο ενώ στο τελευταίο workshop της ομάδας έδωσαν ένα σεμιναριακό μάθημα σε Football analytics.

Ο **L. Egidi** και **Β. Παλάσκας** συνεργάζονται στην ανάπτυξη του λογισμικού "footbayes" (βιβλιοθήκη της στατιστικής γλώσσας προγραμματισμού R).

Ο **L. Egidi, I. Ντζούφρας** και **Β. Παλάσκας** συνεργάζονται στην συγγραφή ενός επιστημονικού άρθρου αξιολόγησης παικτών στο Βόλεϊ.

Ο I. Ντζούφρας και Β. Παλάσκας συνεπιβλέπουν μια διπλωματική εργασία στα πλαίσια του M.Sc. in Statistics του ΟΠΑ και της συνεργασίας με την FSI (Fantasy Sports Interactive)

Η Ομάδα AUEB Sports Analytics

Η ερευνητική ομάδα του Οικονομικού Πανεπιστημίου Αθηνών **AUEB Sports Analytics Group** ιδρύθηκε το 2015 από τους καθηγητές Ιωάννη Ντζούφρα και Δημήτρη Καρλή. Μέλη του είναι σημαντικά μέλη της κοινότητα της αναλυτικής των σπορ όπως οι Leonardo Egidi (Πανεπιστήμιο Trieste), Ιωάννης Κοσμίδης (Warwick), Κωνσταντίνος Πελεχρίνης (Pittsburg), Nial Friel (UCD) και Gianluca Baio (UCL) καθώς επίσης και ο πρώην προπονητής της εθνικής Ελλάδας Βόλεϊ, Σωτήρης Δρίκος και ο νυν προπονητής της Εθνικής ομάδας Μπάσκετ του Κοσόβου, Χρήστος Μαρμαρινός. Η ερευνητική ομάδα είναι υπεύθυνη για της σειρά ετήσιων συνεδρίων με το όνομα AUEB Sports Analytics Workshop (6 συνολικά) ενώ το 2019 διοργάνωσε το διεθνές συνέδριο MathSport 2019 με 200 συμμετέχοντες επιστήμονες από όλο τον κόσμο. Η ομάδα έχει μια σειρά από σημαντικές επιστημονικές δημοσιεύσεις στο χώρο της αναλυτικής των σπορ. Τέλος θα θέλαμε να αναφέρουμε ότι η ομάδα ιδρύθηκε το 2015 λόγω της επίσκεψης του καθηγητή Stefan Kesenne (Πανεπιστήμιο Antwerp & Leuven), σπουδαίου Οικονομολόγου του Αθλητισμού που έπαιξε και ενεργό ρόλο στην υπόθεση Bosman. Ο Stefan Kesenne στήριξε ενεργά την ομάδα μέχρι και το 2021 όπου ξαφνικά απεβίωσε. Η ύπαρξη της ομάδας AUEB Sports Analytics Group οφείλεται σε μεγάλο ποσοστό στη συνδρομή και την έμπνευση που μας έδωσε ο κος Kesenne.