##### Labdarúgás statisztikai adatainak és a mérkőzések kimenetelei közötti összefüggések elemzése a COCO MCM módszerrel, az UEFA statisztikai eredményeinek felhasználásával

Vajon a mindenki által hozzáférhető, publikus adatsorok felhasználásával, egzakt módon meg tudjuk-e mondani, hogy adott vizsgálati időszakban mi lenne a labdarúgó EB eredménye?

Vizsgálatomat a 2012 évi labdarúgó EB-t megelőző időszak UEFA statisztikai adatainak felhasználásával készítettem. A vizsgálat célja, választ keresni arra a kérdésre, vajon megjósolható-e a vizsgált időszak eredményei alapján a bajnokság végeredménye?

Annak érdekében, hogy nyugodt lélekkel tehessünk ilyen jellegű kijelentést, érdemes előtte a rendelkezésre álló statisztikai attribútumokat megvizsgálni, hatnak-e és ha igen, milyen irányban és mértékben a végeredményre. Ehhez a következőkben egy egyszerű metódust mutatok be.

A vizsgálatot exploratív modellezés keretében, COCO Monte Carlo-módszerrel (COCO MCM: <http://miau.gau.hu/myx-free/coco/index.html>) végeztem. A klasszikus MCM egy olyan sztochasztikus szimulációs módszer, amely ál-véletlenszámok generálásának segítségével előállítja egy adott kísérlet valószínűségi modelljét. A COCO MCM azt jelenti, hogy nincs semmilyen korlátozó feltétel megadva a hasonlóságelemzési lépcsőkre (COCO) és ezt a feladatot meg lehet oldani LP-motor nélkül is - az MCM valódi értelmében véletlenszerű találgatással. A COCO MCM modellben elkerülhetjük a tényezők szubjektív irányítását, és a kapott eredményből óvatos következtetéseket lehet levonni.

**Adatgyűjtés (Adatbázis, Kimutatás, Alapadatok, Származtatott (számított) alapadatok)**

A vizsgálatban a fenti kérdésből indultam ki és mindenki által hozzáférhető, hiteles és visszakereshető adatsorokat, adatbázisokat kerestem. A vizsgálat alapját képező adatokat az UEFA honlapjáról töltöttem le. (<http://www.uefa.com)<--pontos> URL?

Igyekeztem a rendelkezésre álló lehető legtöbb attribútumot felvonultatni a 2012–es labdarúgó EB vonatkozásában, a 2009 utáni adatok felhasználásával. Olyan, szinte már közhelyként kezelt mutatókat válogattam, melyek gyakran szerepelnek különféle labdarúgási eredmények statisztikájában, elemzésekben (<http://www.telesport.hu>), illetve a közbeszédben is. Jelen vizsgálatban csak az UEFA adatai kerültek feldolgozásra.

Kialakult a 8 🡨miért pont 8? és miért pont ez a 8?ország csapatát és 10 különböző attribútumot tartalmazó adatbázis. Az adatok mindegyike feldolgozható érték volt (gól, pont, eset, db, százalék), ezért ezek változatlan formában kerültek át a kimutatásba, majd onnan az elemzést tartalmazó alapadat táblába.

Ezzel tulajdonképpen le is zárult az adatgyűjtés fázisa. Már csak az volt hátra, hogy a fenti alapfeltevésből kiindulva potenciális ok-okozati összefüggéseket, illetve ezekre magyarázatokat keressek. Az alaptézisem a végeredményt illetően, a 10 attribútumra kibontva – a laikus focinéző közhelyei alapján - a következőképpen néz ki:

* Elkövetett és elszenvedett szabálytalanságok száma: minél szabálytalanabb egy meccs, annál nagyobb esély van büntető, illetve pontrúgásokra, azaz akció nélküli gólokra, nagy jelentőséggel bír, persze erre alapozni erősen bizonytalan, főleg fair play-t feltételezve. (de mi van a sorsdöntő bírói hibákkal?)
* Kaput eltalált és el nem talált lövések száma: A kaput eltalált lövések száma mindenképpen meghatározó, mert ebből lesznek a gólok (ha ki nem védi a kapus – vagy fokozatos önbizalom-vesztés a csatárokban, akiknek még mindig nem sikerült?), a kaput el nem talált lövések viszont zajnak tekinthetőek. (a kaput el nem talál lövések nem jelenthetnek stresszt az ellenfélnek, mely felőrölheti őket?)
* Labdabirtoklás: közepes jelentőségű, mert zárt védekezésű csapat, kevés labdabirtoklással is lehet eredményes. (ill. taktikai kérdés, be lehet-e rendezkedni eredményes kontrajátékra kevés labdabirtoklással, vagy el lehet-e uralni a teret és az időt a másik csapattól)
* Piros és sárgalapok száma: a hozzájuk kötődő büntető rúgások és kiállítások a gólok számának egyértelmű növekedését kell, hogy mutassák. (vagy éppen a taktikai szabálytalanságok révén egy pici? menekülési esély a gólok elől?)
* Szerzett pontok: ez közepes jelentőségű, a nem elhanyagolható önbizalom erősítését szolgálja. (kellően sok szerzett pont esetén a túlzott önbizalom egy idő után visszaüthet, ill. a sok MÁR megszerzett pont egyre növeli a kiégés veszélyét, esélyét?)
* Szögletek száma: egy kapu közeli akciógól lehetőségével kecsegtetve, nagy jelentőséggel bír. (ha van jó szögletkoncepció és a légtér eluralható a támadók által: vagyis pl. a szögletek száma \* a csapat átlagmagasságával lehetne egy speciálisabb mutatószám?)

**A vizsgálat menete (kvartilis, kvintilis, sextile, dekád, összevonás) és az eredmény**

Az adatbázisból kimutatást készítettem, és ezzel mintegy leszűrtem/tanulási mintává alakítottam a számomra fontos adatokat. Ezeket az adatokat másoltam egy Alapadatok táblába és elkészítettem a tulajdonság- oszlopok szerinti rangsormátrixot, sorszám függvény segítségével, figyelembe véve az rangsor irányultságát (minél nagyobb annál jobb vagy minél kisebb annál jobb). Az így előállt rangsormátrixot az Y komponens adataival kiegészítve, a COCO online additív standard verzió alkalmazásával (vagyis az Y-ra vonatkozó termelési függvények kialakításával) igyekeztem elemezni. Sajnos az inverzióval történő hitelesség ellenőrzésen rendre megbuktak a próbálkozásaim annak ellenére, hogy újabb és újabb attribútumokat vontam be az elemzésbe, bővítve az adatsort.

Ezeket a COCO-STD-lépéseket pontosabban kellene dokumentálni az XLS-ben is és itt is…

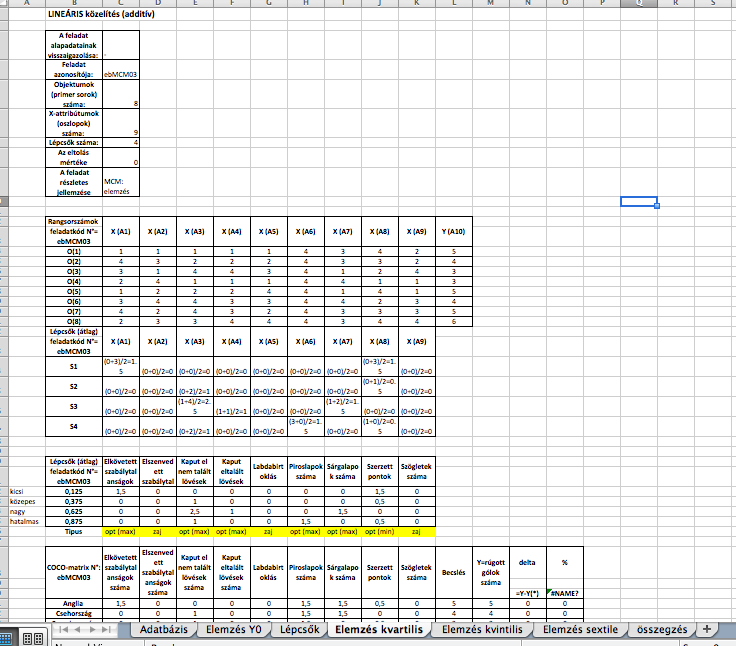
Az így tapasztalt sikertelenség után tértem át a Monte Carlo módszer használatára, mely nagyobb szabadságot engedett az Xi és az Y közötti összefüggések leképezésében. Ugyan itt is el kellett készíteni az Alapadatok táblát, és a rangsor mátrixokat, de az irányultságot nem kellett megadni (helyesebben mindenhol alapértelmezésben a minél nagyobb annál jobb beállításból kellett kiindulni a későbbi értelmezések megkönnyítése érdekében), ami már egy szubjektivitási hibalehetőséget kiküszöbölt. A rangsormátrix(ok) elkészítése viszont eltérő módon történt, mégpedig a lépcsők számától függő mennyiségben. Ezeket külön-külön futtattam, melyek későbbi összevonásával finomítani lehetett a kapott végeredményt.

Először meghatároztam a lépcsők számát, ami azt jelenti, hogy az adatokat méretük alapján ilyen kevés csoportba (4-5-6) kellett illetve érdemes besorolni. A legegyszerűbb a 3 lépcsős modell. A három lépcső adata grafikusan elképzelve vagy egy egyenesbe rendeződik, vagy valamilyen egyszerű görbét ír le. Az egyenes lefutását tekintve vagy nő, vagy csökken, vagy stagnál, a görbe pedig vagy minimummal, vagy maximummal rendelkezik, tehát mindegyik egyszerűen értelmezhető. Az első lépcső a kicsi, a második lépcső a közepes és a harmadik lépcső a magas hatótényezővel bíró adatokat tartalmazza.

Ettől felfelé haladva a lépcsők hullámzani kezdhetenek, fellép a polinomhatás és nő a bizonytalanság a következtetések levonása kapcsán. Annak érdekében, hogy kellően részletes képet kapjunk az elemzés során, nem feltétlenül elegendő a 3 lépcsős modell elkészítése. Készült ilyen??? Mi lett az eredménye? Hogyan értelmezzük a tanulás pontosságát?

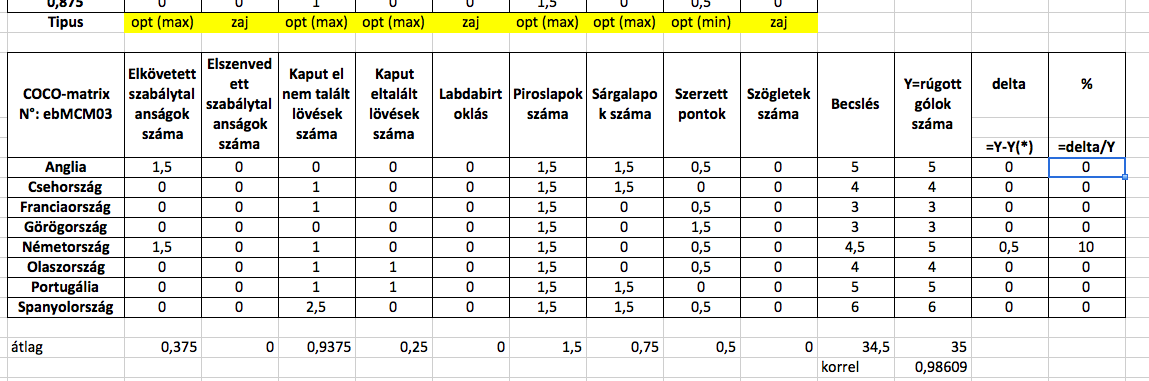
A besorolás és a rangsormátrix elkészítése úgy történt, hogy az adott attribútum minimum és maximum értéke által leírt szakaszt a lépcsők számának megfelelően egyenlő részekre osztva, az adott szakaszba tartozó értékek (a szakaszhatárokhoz kapcsolt „HA” függvények segítségével) kaptak azonos rangsorszámot. Ezt követően történhetett meg a rangsormátrix feltöltése és a futtatás. Továbbfejlesztési lehetőségként merül fel, hogy a szakaszhatárok levezetése is legyen a módszer része!

A vizsgálatban először négy lépcsőben, másodszor öt lépcsőben, végül hat lépcsőbe tagozva vizsgáltam az adatokat a COCO-MCM modellben. A modell lépcsői a hatás mértékét mutatják meg számunkra, mely az egyes lépcsőszámhoz tartozó elemző munkalapokon szépen nyomon követhető.



. ábra Kvartilis adatok az MCM módszerrel (forrás: saját ábrázolás)🡨ellenben a sárga cellák tartalma biztos, hogy helyes? Mi a szabályrendszer? Mik a lehetséges szövegsablonok? Ha a lépcsők értéke ilyen vagy olyan, akkor ….<--hány féle állapot lehetséges következményként? pl. monoton növekedés, csökkenés, telítési vagy platójellegű összefüggés, stb…

Az így kapott eredményeknek két vetülete is van. Egyrészt az 1. ábrán látható módon attól függően, hogy az adott tényező melyik lépcsőben kap magas értéket, annyira meghatározó az eredmény alakításában. Az első lépcső a legkisebb, míg az utolsó a legnagyobb bemeneti jelszintet jelenti. Másrészt, az eredmények visszavetíthetőek az eredeti rangsormátrixra, ahol az adott lépcsőbe tartozó cellák kapják ezeket az értékeket. (2. ábra)



. ábra Országokra/csapatokra visszavetített hatásadatok a kvartilis elemzés eredménye alapján (forrás:…)

🡨helyes-e az a modell, melyben a tények és a becslések összege nem azonos?

A 2. ábra magyarázata tehát, hogy például Görögország szerzett pontjaira leginkább a rúgott gólok száma, legkevésbé pedig a kaput el nem talált lövések száma gyakorolt hatást. Az elkövetett szabálytalanságok száma közepes erővel bírtak.

Hogy is van ez? Az Y = rúgott gólok száma! Tehát hol van itt a görög szerzett pont???

A görög rúgott gólok felhajtó erejét a piros lapok és a szerzett pontok adják?!

Az előző fejezetben részletezett feltevéseimhez képet, érdekes kép rajzolódott ki, már a négylépcsős modellben is. Eszerint a végeredmény szempontjából teljesen irrelevánsnak tekinthető az elszenvedett szabálytalanságok száma, a labdabirtoklás mértéke és a szögletek száma. Elhanyagolható jelentőségű az elkövetett szabálytalanságok és a szerzett pontok száma. Érdekes módon nagy szerepet játszik a kaput eltalált, illetve el nem talált lövések és a sárgalapok száma, míg döntő súlyú a piros lapok száma.

Az ötlépcsős modell, a nagyobb felbontás okán változások sorát hozta. Bár a zajnak tekinthető tényezők száma három maradt, de ez meglepő módon a kaput eltalált lövések, a szerzett pontok és a szögletek száma lett. Maradt kis jelentőségű az elkövetett szabálytalanságok száma és társult hozzá a korábban zaj kategóriában szereplő, elszenvedett szabálytalanságok száma, tehát szám szerint ez a lépcső is kéttagú maradt. Előfeltevésemet beigazolni látszó, közepes jelentőséggel bírva megjelent a labdabirtoklás aránya és további két polinomizálódó tényező (kis jelentőséggel) mégpedig a sárgalapok száma és a kaput el nem talált lövések száma. Ami nem változott, hogy a piros lapok száma döntő jelentőségű maradt.

Hogyan, honnan olvasható le, hogy valami kishatású, zaj, stb. (vö. előző kérdés a szövegsablonokról és ezek HA()-függvényeiről…)

A fentiek fényében el kellett készíteni a hatlépcsős modellt is, hogy igazságot tegyen a fenti rendezetlenségben. A zaj csoport maradt három tagú, szögletek száma, kaput eltalált lövések száma és visszatért az elszenvedett szabálytalanságok száma. A kicsi jelentőségű lépcső szintén két tagú maradt, az elkövetett szabálytalanságokkal és a kaput el nem talált lövések számával. A közepesnél kicsit nagyobb súllyal tűnik fel a szerzett pontok száma, míg a két polinom megmaradt, csupán a labdabirtoklás jelentősége nőtt. A piros lapok száma viszont továbbra is döntő jelentőségű maradt.

Mivel a polinomok száma nem nőtt és az eredmények közel azonosak voltak, nem láttam értelmét újabb, többlépcsős modell elkészítésének. Így már csak az maradt, hogy össze kellett vonni a különböző MCM-rétegeredményeket. Az összevonás lényege, hogy az általuk leírt szakaszokat 1-nek tekintjük. A lépcsők száma szerinti határoló pontokat felírtam a képviselt intervallum középértékével, majd melléjük írtam a lépcsők értékeit attribútumonként.

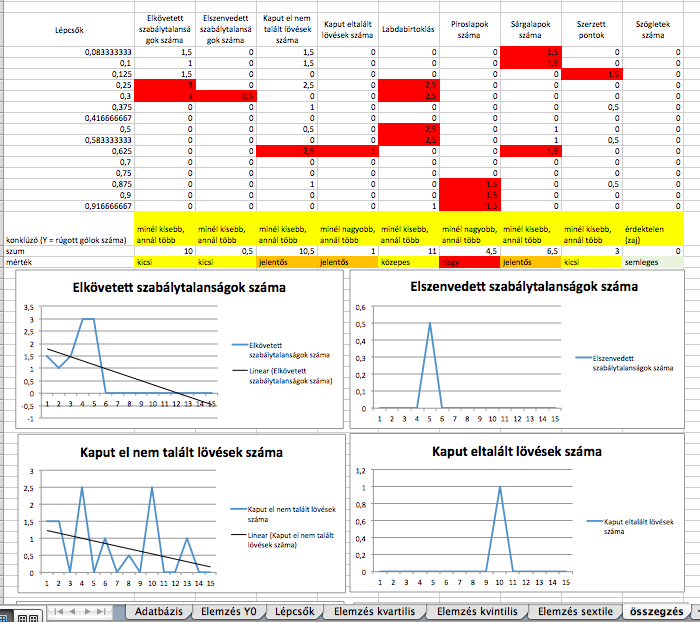
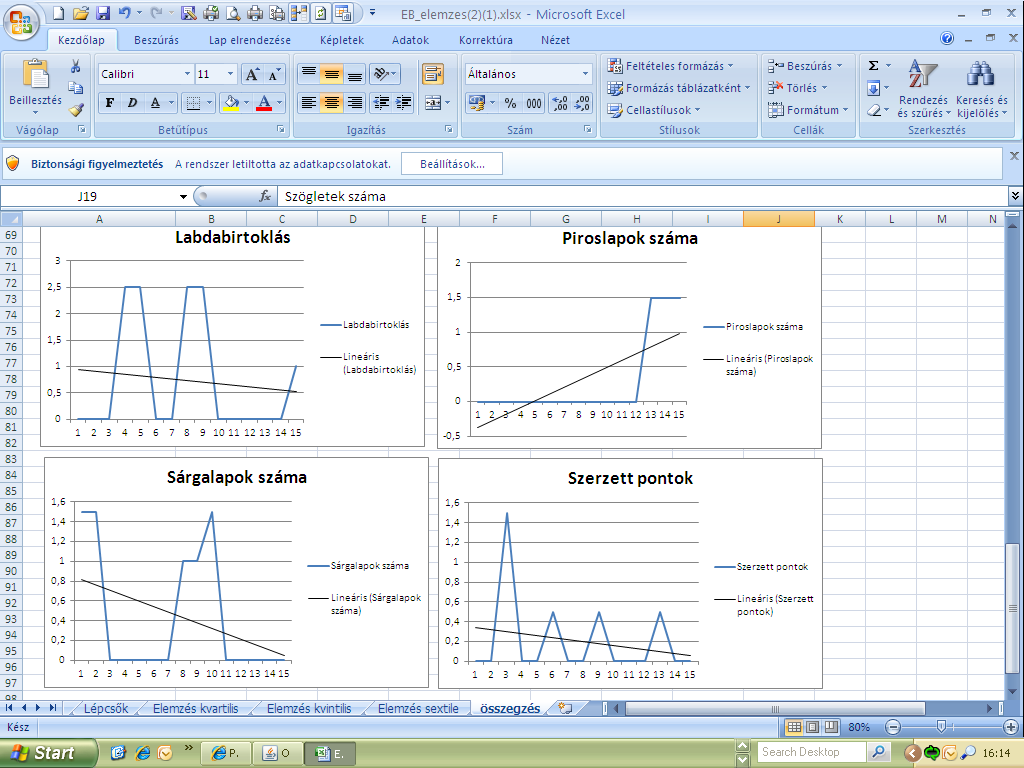
A határoló pontokat nagyság szerint sorba rendezve, kaptam azt az összképet, amit a rétegek mindösszesen sugallnak. Az ezekhez illesztett ábrák, valamint a hozzájuk rendelt trendfüggvények alapján már lehet feltételezni, milyen is a ceteris paribus alakzat a háttérben. (3. ábra)

**Értékelés, következtetés, javaslatok**

Az exploratív modellek által kiszámított lépcsőket összevonva, az egyes táblázatok és az aggregált lépcsők táblázata alapján próbáltam következtetéseket levonni a modellből. Minden végkövetkeztetés beszédes háttere a 3. ábra, melynek grafikonjait a szemléletesség kedvéért trendfüggvényekkel egészítettem ki.

Ez alapján, előfeltevésem mintegy cáfolataként egyértelműen zaj a szögletek száma, ahol egyetlen modell-nézet sem tudott magyarázó erőt vélelmezni (vö. végig nulla hatás a rúgott gólok számára), így a végeredmény tekintetében elhanyagolható. Érdekes, objektív ellenőrzési lehetőség lenne, vajon hány gól esett a statisztikák hátterében közvetlenül szöglet nyomán az összes gólhoz képest?

Kicsi (🡨mitől kicsi a kicsi jelentős?) jelentőséggel bír három tényező: az elkövetett és elszenvedett szabálytalanságok, illetve a szerzett pontok száma. Érdekes módon fordított arányosságban áll mindhárom a végeredmény kimenetelével. Tehát a fenti három tényező értéke minél kisebb, minél kevesebb szabálytalanság történik és minél kevesebb a szerzett pontok száma, annál jelentősebb lesz a végeredmény. Viszont ezek befolyásoló hatása kicsi, de – a szabálytalanságok esetén a görbe lefutása alapján – egyértelmű. A polinomként viselkedő szerzett pontok száma, az előfeltevést bizonyítja, a végeredményre gyakorolt bizonytalan, de kicsi hatásával. Vajon lehetne itt valamilyen módon a hatások szignifikáns jellegét kimondani?



. ábra Aggregált lépcsők függvény és a tulajdonságok diagramjai trendfüggvényekkel (forrás: …)

Az egyedüli közepes (🡨mitől közepes a közepes?) jelentőséggel bíró tényező a labdabirtoklás mértéke, mely fordított arányban áll a végeredménnyel és polinom. Ez is az előfeltevés malmára hajtja a vizet (miért is? mi is volt a pontos szövege az előfelvetésnek??? volt egyáltalán pontos előfelvetés?), hiszen gólt valóban az tud rúgni, akinél a labda van, de extrém módon kisarkítva, elég egy labdába érés is, ha abból éppen a győztes gól születik meg. Sőt úgy tűnik a fordított arányossági viszony miatt, hogy az optimális helyzet, ha a labdabirtoklás mértéke kicsi. Klasszikus történelmi példák a labda túlbirtoklásra: ???

Az Y változónk szempontjából jelentős súllyal ható három tényező, a kaput eltalált, el nem talált lövések száma és a sárga lapok száma már változatosabb és némi magyarázatra szorul. A kaput eltalált lövések száma a görbe lefutása alapján is egyértelmű, pozitív hatást gyakorol a végeredményre, egyenes arányosságban állva azzal. Nyilvánvalóan a döntő jelentősége azért nem valósulhat meg, mert kapus is van a kapuban, tehát a lövések egy része nem ér el gólt. A másik két tényező viszont fordítottan arányos az Y-nal és ráadásul polinomként viselkedik. A hullámzás és bizonytalanság oka feltehetően az előfeltevésben is jelzett ok, hogy a sárgalapok száma és annak hatása egy meccsen bizonytalan. Pontosan ugyan ez a helyzet a kaput el nem talált lövésekkel, hiszen például egy kapufáról vagy kapusról kipattanó lövés nem mindig kerül értékesítésre. A fordított arányosságot is figyelembe véve minél kevesebb rúgás megy mellé és minél kevesebb sárga lap gyűlik össze, annál több a rúgott gól. Ha egybe olvassuk a szabálytalanságok számánál tapasztaltakat a sárga és piros lapok potenciális hatásaival, akkor kijelenthető vajon, hogy akkor esik több gól, ha szabályosabb a mérkőzés? Ha igen, akkor a destruktív játékszemlélet szemmel látható térnyerése legitimálható? S végül, ha megéri szabálytalankodni, nem kellene súlyosabban büntetni a taktikai beavatkozásokat? (pl. automatikus kiállítások lapok helyett?) Egy ilyen elemzésnek lehet stratégiai hatása a játékszabályok alakulására? Különösen, ha cél a tömegek lekötése révén (=több gól) a játék marketingerejének növelése/stabilizálása?

Végül, de nem utolsó sorban a piros lapok száma, egyértelműen döntő hatással van a kimenetel eredményére, mind az adatok, mind a görbe futása alapján. Nos, a sárga lapok számának bizonytalan, de jelentős hatása mellett ez egy érdekes jelenség, viszont érthető, mert ez az adott csapat létszámát csökkenti és előbb-vagy utóbb gólt fog eredményezni feltéve, ha a másik csapat élni tud az emberelőnnyel. (vö. előző kommentárok!)

Vajon a polinomizálódás (hullámzás) milyen mértéke mellett illik a trend meredekségét nem szignifikánsnak (azaz értelmezésre nem méltónak) tekinteni?

Összefoglalva tehát, a rúgott gólok számát egyértelműen befolyásolja a piros lapok száma minél több a piros lap, annál nagyobb a rúgott gólok száma. Szintén igaz, hogy minél több rúgás találja el a kaput, annál több a rúgott gól, és ugyan ebben az irányban hat, ha minél kevesebb a sárga lap és a kaput el nem talált lövés (ami ugye a kaput eltalált lövések logikus inverze). Figyelembe vehető még a labdabirtoklás mértéke, de jelentősége már a vagy-vagy kategória, ráadásul a több rúgott gólért, kevesebbet kell birtokolni a labdát. A végeredmény alakulásában ugyan még beleszól az elkövetett vagy elszenvedett szabálytalanságok száma, illetve a szerzett pontok mennyisége, de már csak elhanyagolható mértékben és minél kevesebb van belőlük, annál jobb lesz a végeredmény. Végül meglepő módon teljesen figyelmen kívül hagyható a szögletek száma.

A felvetett kérdés immár „statisztikai” támaszt is kapott. Teljesen leegyszerűsítve, és gyakorlatiasan megfogalmazva, az alábbiakban látom vizsgálódásom értelmét és hasznát. Amennyiben focidrukker vagyok és rendelkezem egy folyamatosan frissülő adatbázissal - mely tartalmazza az adott csapat(ok) által elkövetett és elszenvedett szabálytalanság(ok) számát, kaput eltalált és el nem talált lövés(ek) számát, labdabirtoklás arányát, piros és sárga lap(ok) számát, szerzett pontokat és szögletrúgások számát - akkor meg tudom-e mondani, és ha igen, milyen eséllyel, hogy a kedvenc csapatom hol végez az adott bajnokságban. Innen viszont már csak egy ugrás, hogy az alkalmazás egy sportfogadási oldal – annak adatbázisát használva - virtuális „jósló állatkája” legyen, ahogyan ezt a média oly nagy szeretettel fel is karolta az utóbbi időkben.

Milyen eséllyel lehet nyerni a sportfogadásban ezek után?

Nem inkább az edzők és a FIFA számára tanulságos ez az elemzési pontosság?

# Források

Mi az EB?

<http://hu.wikipedia.org/wiki/Labdarúgó-Európa-bajnokság>

A 2012-s EB adatai

<http://hu.wikipedia.org/wiki/2012-es_labdarúgó-Európa-bajnokság>

+1 kép

