5. fejezet

Vizuális analízis

Az adatvizualizáció, akárcsak a matematikai statisztika, napjaink természet- és társadalomtudományainak és mérnöki gyakorlatának alapvető eszköze. Fejlődése összefonódik alkalmazási területeinek fejlődésével; már a 10. századból ismert olyan ábra, mely hét égitest elhelyezkedésének változását demonstrálja térben és időben – mai szóhasználatunkban idősorként [52]. A 19. század első felére a (statikus) statisztikai grafika alapvető eszközeinek többsége kialakult: így például az oszlopdiagram (*bar chart*[1]), a hisztogram, az idősorábra (*time series plot*), a szintvonal-ábra (*contour plot*) vagy a szórásdiagram (*scatterplot*) [51]. A 70-es évektől kezdve az adatvizualizáció mai gyakorlatához vezető fejlődési folyamatot alapvetően befolyásolta a felderítő adatanalízis (*exploratory data analysis*), mint önálló statisztikai diszciplína kialakulása és a számítógéppel megvalósított adatábrázolás lehetővé, majd gyakorlatilag egyeduralkodóvá válása. Bár az egyes szakterületek szükségletei még ma is életre hívnak újabb és újabb diagram-típusokat (melyek sokszor csak az ismert típusok variációi), a modern statisztikai adatvizualizáció legfontosabb minőségi újításai a magas dimenziójú statisztikai adatok kezelése, valamint az interaktív és dinamikus megjelenítési technikák. Kialakulóban vannak, de messze nem kiforrottak azok az általános vizualizációs módszerek, melyek segítségével extrém méretű adatkészletek – divatos szóhasználatban: „Big Data" problémák – is célszerűen szemléltethetővé válnak (lásd pl. [120]).

Az adatvizualizáció és a segítségével megvalósított vizuális analízis rendkívül szerteágazó témák; fejezetünk célja a sokváltozós statisztikához kapcsolódó alapvető és általános vizualizációs technikák és az ezekre épülő elemzési megközelítések ismertetése. Alkalmazási útmutatóként – különösen a statikus diagramok tekintetében – az `R` [91] nyílt forráskódú és ingyenes statisztikai számítási környezetben rendelkezésre álló megoldásokra fogunk hivatkozni; meg kell azonban említenünk, hogy a diagramok többségét ma már mindegyik meghatározó adatelemzési eszköz támogatja.

[1] A jellemzően angol nyelvű, nemzetközileg elfogadott terminológia helyett a fejezetben törekszünk a magyar megfelelők használatára – az eredeti megjelölésével. Felhívjuk azonban az olvasó figyelmét arra, hogy a terület számos szakkifejezésének nincs egyértelműen és általánosan elfogadott magyar fordítása.

5.1. Felderítés, megerősítés és szemléltetés

Bár az adatvizualizáció szinte minden adatelemzési feladatban megjelenik, súlya és főként alkalmazásának módja az adatelemzés célja és fázisa szerint más és más. A vizualizáció eszköze lehet *a*) az elemzendő adatok *megértésének* és *hipotézisek megsejtésének*; *b*) hipotézisek és modellek megerősítésének, vagy *c*) az eredmények – lehetőleg minél szemléletesebb – *prezentálásának*. Az első két tevékenység-kategóriára szokásosan mint *felderítő* adatelemzés (*Exploratory Data Analysis – EDA*), illetve *megerősítő* adatelemzés (*Confirmatory Data Analysis – CDA*) hivatkozunk.

A John W. Tukey amerikai matematikus és statisztikus munkássága által megalapozott EDA leginkább egy „statisztikai tradíció", melyet röviden a következőképp jellemezhetünk [9].

1. Az adatelemzési folyamat nagy hangsúlyt fektet az adatok általánosságban „megértésére" - az adatok által leírt rendszerről, az adatokon belüli és az adatok és általuk leírt rendszer közötti összefüggésekről minél teljesebb fogalmi kép kialakítására.

2. Az adatok grafikus reprezentációja, mint ennek eszköze, kiemelt szerepet kap.

3. A modelljelölt-építés és a hipotézisek felállítása iteratív módon történik, a modell-specifikáció – reziduum-analízis – modell-újraspecifikáció lépéssorozat ismétlésével.

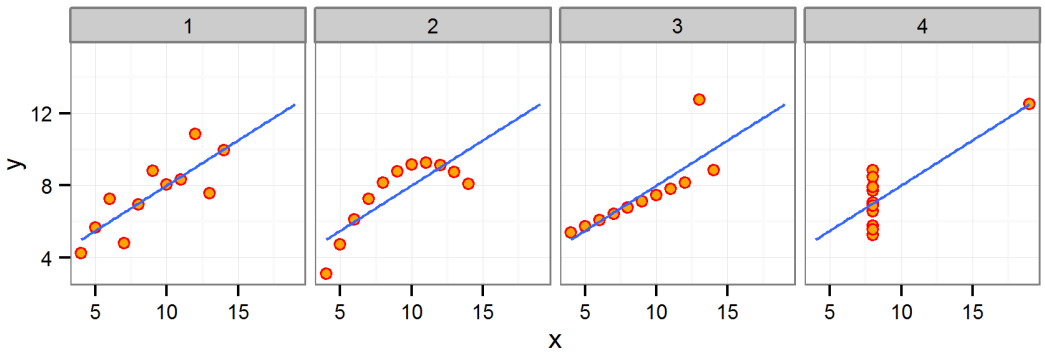
4. A CDA-val összevetve előnyt élveznek a robusztus statisztikai mértékek és a részhalmazok analízise.

5. Az EDA „detektívmunka" jellege miatt nem szabad bevett módszerekhez dogmatikusan ragaszkodni; a folyamatot köztes sejtéseink mentén flexibilisen kell alakítani, előfeltételezéseinket (pl. hogy két változó között korreláció „szokott" fennállni) megfelelő szkepszissel kezelve.

Természetéből adódóan az EDA legtöbbször egy erősen ad-hoc folyamat; jellemezhető úgy is, mint az adatok (jellemzően) alacsony dimenziószámú grafikus vetületeinek intuíció- és szakterület-specifikus tudás által vezérelt bejárása addig, amíg klasszikus statisztikai elemzést érdemlő hipotézisekre nem jutunk.

A megfelelő vizualizáción keresztül összefüggések megsejtésének iskolapéldája Dr. John Snow története. Dr. Snow 1854-ben egy londoni kolerajárvány alkalmával egy pontozott térképet használva a halálesetek vizualizálására felfedezte, hogy a járvány oka egy fertőzött kút a Broad Streeten; a kút nyelét leszereltetve a járvány megszűnt. (A történet második része valószínűleg nem igaz; lásd [86].) A CDA túl korai, EDA-t nélkülöző alkalmazásának veszélyeire a klasszikus intő példa pedig „Anscombe négyese" (*Anscombe's quartet*)[5]: négy kétváltozós adatkészlet, melyeknek ugyan átlaga, varianciája, korrelációja és regressziós egyenese – azaz klasszikus statisztikai jellemzői – megegyeznek, mégis minőségileg különböző összefüggéseket rejtenek (5.1. ábra)[2].

[2]Anscombe négyese az `R` beépített, `anscombe` néven előhívható adatkészlete.



5.1. ábra. Anscombe négyese szórásdiagramokkal vizualizálva

Az alfejezet további részében bemutatjuk a vizuális EDA során leggyakrabban alkalmazott diagram-típusokat. Bár ezeket sokszor alkalmazzuk a CDA támogatására is, ott jellemzően modellspecifikus vizualizációra van szükség (pl. lineáris regresszió reziduumainak vizsgálata).

5.2. Egydimenziós diagramok

Egy sokváltozós megfigyeléskészlet változóinak peremeloszlás-vizsgálata az EDA első lépései között szokott szerepelni. Amennyiben a változó kategorikus, úgy a közismert oszlopdiagram és változatai szolgálhatnak a kategóriák számosságának vizualizációjára.

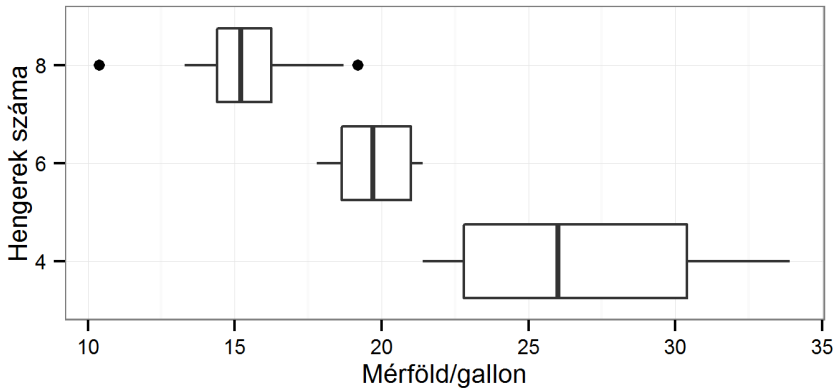
Az egy, folytonos változóra vonatkozó megfigyeléseket reprezentálhatjuk közvetlenül egy tengelyen, ún. pontdiagramon (*dot plot*) - a szórásdiagram egyfajta „leskálázásaként". Az egydimenziós, folytonos esetben azonban jellemzően nem közvetlenül a megfigyeléseket, hanem vagy meghatározó leíró statisztikáikat, vagy eloszlásukat szeretnénk vizualizálni. Mindkét esetben jellemző – legalább a felderítő analízis során – hogy *nemparaméteres* technikák alkalmazására törekszünk, azaz a változó eloszlásával kapcsolatban nem kívánunk előfeltételezésekkel élni.

5.2.1. A doboz-ábra

Egy egyváltozós megfigyelés-sokaság alapvető leíró statisztikáinak legnagyobb kifejezőerejű vizualizációs eszköze a doboz-ábra (*box plot*) és különböző variációi [85]. A doboz-ábra öt alapvető jellemzőt szemléltet: a minimumot, az első kvartilist, a mediánt, a harmadik kvartilist[3] és a maximumot. A kvartilisek távolságát, mint az eloszlás „középső felét", egy

[3]Az első és a harmadik kvartilis helyett precízebb lenne alsó és felső „sarkalatos pontról" – *hinge* – beszélnünk. Az alsó sarokpont definíció szerint az $([(n+1)/2]+1)/2$-ed rendű statisztika, mely ha nem egész, akkor a szomszédos statisztikák átlagát használjuk; a gyakorlatban ez azonban jó közelítéssel az első kvartilis. A felső sarokpont hasonlóan definiálható.

„doboz” reprezentálja; az első kvartilis alatti és a harmadik feletti részt jellemzően egy vonal, „bajusz” (*whisker*). Egy jellemző variáció, hogy ez a vonal nem a minimumig és a maximumig nyúlik, hanem legfeljebb a kvartilisek közötti távolság (*Inter-Quartile Range*, $IQR = Q_3 - Q_1$) 1,5-szereséig — az ezen kívül eső megfigyeléseket pontként ábrázolva. Emellett – bár a doboz-ábrát egydimenziós vizualizációs technikaként a legegyszerűbb bevezetni – a gyakorlatban legtöbbször a megfigyeléseket egy kategorikus változó mentén részhalmazokra bontva használjuk. Az 5.2. ábra példa doboz-diagramja is kétdimenziós ebben az értelemben[4].



5.2. ábra. Példa doboz-ábrára: személygépjárművek üzemanyag-hatékonysága

Egy doboz-diagram önmagában egy egyváltozós adatkészlet centrális tendenciája, diszperziója és ferdesége (*skewness*) felmérésének praktikus eszköze. Több eloszlás esetén (pl. kategorikus változó mentén bontás mellett) ezen jellemzők gyors, vizuális összehasonlítására is alkalmas. Hátránya, hogy az eloszlást nagyon erősen absztrahálja; így például a multimodalitás jellemzően nem olvasható le róla.

5.2.2. Hisztogram

Legyen $x \in \mathbb{R}$ változó, melyen az $a, b \in \mathbb{R}$ intervallumon n megfigyeléssel rendelkezünk. Képezzük $[a, b)$ egy L nemátfedő intervallumokból (*bin*-ekből, cellákból) álló partícionálását: $T_l = [t_{n,l}, t_{n,l+1}), l = 0, 1, 2, \ldots, L-1$, ahol $a = t_{n,0} < t_{n,1} < t_{n,2} < \ldots < t_{n,L} = b$. Legyen I_{T_l} az l-ik cella indikátor-függvénye és legyen $N_l = \sum_{i=1}^{n} I_{T_l}(x_i)$ a T_l-be eső minták száma ($l = 0, 1, 2, \ldots, L-1$, $\sum_{l=0}^{L-1} N_l = n$). Ekkor a hisztogram, mint a változó eloszlásának becslője, a következőképp definiálható ([69], 80. oldal):

$$\hat{p}(x) = \sum_{l=0}^{L-1} \frac{N_l/n}{t_{n,l+1} - t_{n,l}} I_{T_l}(x).$$

[4]Az ábra az `R` beépített `mtcars` adatkészlete felett készült, mely 32, 1973–74-es személygépkocsi-modell tíz aspektusát írja le.

A gyakorlatban általában azonos, de a minták száma által befolyásolt cellaszélességet használunk ($h_n = t_{n,l+1} - t_{n,l}, l = 0, 1, 2, \ldots, L-1$). A struktúra vizualizációja közismert; az 5.7. ábrán láthatunk két példát. A hisztogram az EDA szempontjából legfontosabb hátránya, hogy alakja erősen érzékeny mind az első cella kezdőpontja, mind pedig a cellaszélesség megválasztására. Becslőként is komoly hiányosságai vannak; numerikus sűrűségbecslésre legtöbbször alkalmasabbak a kernel-sűrűségbecslők (lásd pl. [69], 4.5. fejezet). Vizuális elemzés során azonban egyszerűbb, az „adatokhoz közelebb eső" interpretálhatóságuk miatt mégis a hisztogramok előnyben részesítése javasolt.

5.3. Kétdimenziós diagramok

Statikus vizualizáció esetén a két változó közötti interakciót szemléltető, általános célú diagramok szempontjából a két kategorikus változó esete érdemel kiemelt figyelmet. A folytonos-folytonos esetben alkalmazható szórásdiagramok és hőtérképek (*heat map* – valójában 2D hisztogramok) közismertek; a kategorikus-folytonos esetben alkalmazhatunk pl. kategóriánként alkalmasan színezett hisztogramokat (lásd pl. később az 5.8. ábrán) vagy a már bemutatott kondicionált doboz-diagramokat.

A kétváltozós kategorikus-kategorikus esetben elsődleges célunk a kategória-kombinációk relatív számosságának felmérése lehet. Erre általában a mozaik diagram (*mosaic plot*) vagy a fluktuációs diagram (*fluctuation plot*) a legalkalmasabbak. Ezek azonban n-dimenziós diagramtípusok, melyeknek a két változó megjelenítése speciális esete.

5.4. n-dimenziós diagramok

Kettőnél több változó megjelenítésére két alapvető diagramtípust mutatunk be: a tisztán kategorikus esetre a mozaik diagramot (és a variánsának tekinthető fluktuációs diagramot), a tisztán folytonos esetre pedig a párhuzamos koordináta (*parallel coordinates*) diagramot. A párhuzamos koordináta diagramon kategorikus változók is megjeleníthetőek a kategóriákhoz számérték rendelésével, azonban mint látni fogjuk, ez jellemzően csak akkor eredményezhet „jól olvasható" diagramot, ha a kategorikus változónak megfelelő tengelyek nem szomszédosak.

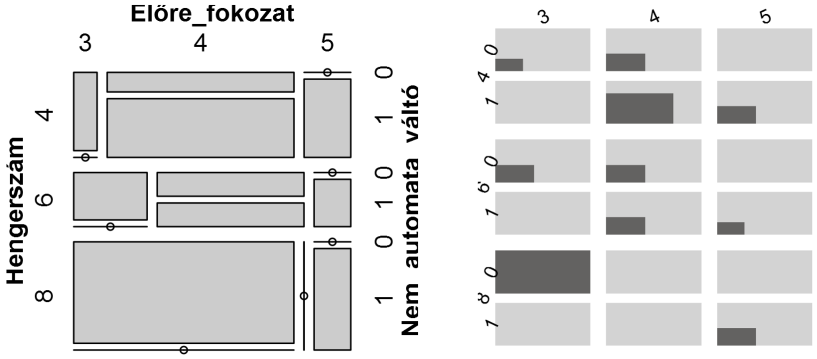
Megjegyezzük, hogy a folytonos esetben elterjedten használt még a szórásdiagrammátrix (*scatterplot matrix*, *SPLOM*); hogy egy sokváltozós adatkészlet változó-vektorát megfeleltetjük egy mátrix sorainak és oszlopainak, a cellákban pedig a megfelelő változópárok szórásdiagramját helyezzük el. A SPLOM és variánsai – a párhuzamos koordinátákkal ellentétben – magasabb változószámnál korlátozott használhatósága miatt mélyebb bemutatásuktól eltekintünk.

5.4.1. Mozaik és fluktuációs diagram

A mozaik diagram kategorikus változók érték-kombinációi előfordulási gyakoriságainak területarányos vizualizációja. Az ábrát alkotó „csempéket" vagy „lapokat" (*tiles*) egy négy-

zet rekurzív vízszintes és függőleges darabolásával kapjuk. Az 5.3. ábrán látható példa az `mtcars` adatkészlet három változóját helyezi el mozaik diagramon, hengerek száma, előre fokozatok száma és a váltó nem automata volta sorrendben. Egy negyedik változó az előre fokozatok száma alá eső, annak értékeit rendre saját lehetséges értékeivel aláosztó faktorként jelenne meg. Egy ötödik változó a hengerszám – nem automata váltó bontást finomítaná tovább, és így tovább. A mozaik diagramok effektív olvashatósága természetesen adatfüggő is, de elmondható, hogy körülbelül 8 változónál többet általában semmiképp sem érdemes alkalmazni. Mindemellett az olvashatósággal már 4–5 változó esetén adódhatnak problémák.

A fluktuációs diagram a mozaik diagram magasabb dimenziószámnál nehezen olvashatóságát próbálja orvosolni. Az egyes érték-kombinációkhoz azonos méretű lapokat rendelünk, ezáltal a kombinációk könnyen beazonosíthatóvá válnak. A legnagyobb elemszámú lapot teljesen kitöltjük, a többit pedig az előfordulások relatív gyakorisága alapján területarányosan. (A kitöltő téglalapok oldalaránya a mozaik diagramok rekurzív bontásával szemben egységesen ugyanaz; lásd 5.3. ábra.) Hátránya, hogy a kitöltő idom nem hordozza közvetlenül az egy-egy változóban értelmezett relatív gyakoriságot vizuális információként.

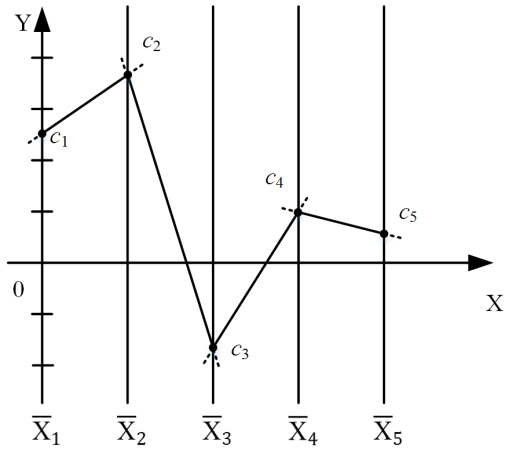


5.3. ábra. Mozaik diagram és fluktuációs diagram

5.4.2. A párhuzamos koordináta ábra

A párhuzamos koordináta diagram [68], mint a sokváltozós relációk vizualizációjának eszköze, egy igen egyszerű ötleten alapul. A klasszikus Descartes-féle ortonormált koordinátarendszer „gyorsan kimeríti a síkot"; már három dimenzió esetén is projekciókra van szükségünk. Ennek oka az, hogy a tengelyek merőlegesek (de legalábbis szöget zárnak be). A párhuzamos koordináták ezzel szemben a szokásos Descartes-féle koordinátákkal ellátott $\mathbb{R}^2$ euklideszi síkban egymástól azonos (egységnyi) távolságra elhelyezi az $\mathbb{R}$ valós egyenes N másolatát, és ezeket használja tengelyként az $\mathbb{R}^N$ N-dimenziós euklideszi tér pontjainak ábrázolására. A $(c_1, c_2, \ldots, c_n)$ koordinátákkal rendelkező $C \in \mathbb{R}^N$ pont képe az a teljes poligon-vonal, melynek N darab, a szegmenseket meghatározó pontjai rendre a

párhuzamos tengelyekre eső $(i-1, c_i)$ pontok $(i = 1, \ldots, N)$. Ezt a kölcsönösen egyértelmű leképezést szemlélteti az 5.4. ábra[5].



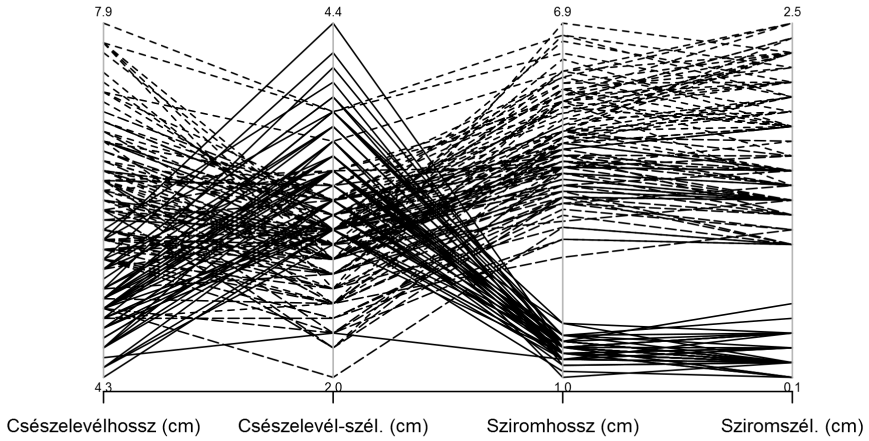
5.4. ábra. $\mathbb{R}^N$-beli pont ábrázolása párhuzamos koordinátákkal

Figyeljük meg, hogy a módszer „helyigénye" lineárisan skálázódik a változók számában, szemben pl. a szórásdiagram mátrix négyzetes növekedésével. Így a párhuzamos koordináta diagram magas dimenziószámú adatkészletek ponthalmazainak projekció nélküli *áttekintésére* különösen alkalmas eszköz. Az 5.5. ábrán a híres Fisher-féle írisz-adatkészlet[6] párhuzamos koordinátákra leképezése látható. Példánk egyben azt is jól szemlélteti, hogy miért kisebb a párhuzamos koordináták jelentősége a statikus vizualizáció területén, mint az interaktív technikáknál: nagyszámú pont esetén az ábra gyorsan átláthatatlanná válik, különösen faktorváltozók szerinti megkülönböztető színezés nélkül. Amennyiben azonban rendelkezésre állnak a megfelelő interakciók – mint pl. részhalmaz kiválasztása, tengelyen intervallum kiválasztása –, a párhuzamos koordináták különösen hatékony EDA eszközt adnak kezünkbe.

Ennek fő oka, hogy számos síkbeli és térbeli alakzat a párhuzamos koordinátákkal ábrázolás során jellegzetes mintává képződik le, így az EDA felfogható egyfajta mintafelismerési problémaként. A talán legegyszerűbb eset ennek szemléltetésére a pont-vonal dualitás. Mint láttuk, egy pont két, szomszédos párhuzamos koordináta-tengelyen értelmezett képe egy szakasz. Az ezen két tengelyhez tartozó koordináták síkjában felvett egyenes képe viszont párhuzamos koordinátákban egy *pont* abban az értelemben, hogy az egyenes pontjai leképezésének tekinthető szakaszok egy pontban fogják metszeni egymást. (Nem feltétlenül a két párhuzamos tengely között.) Ez azt jelenti, hogy amennyiben egy két tengely közötti szakaszsereg egy pontban metszi egymást egy párhuzamos koordináta ábrán, úgy a tengelyek (Descartes) koordinátáinak síkjában egy egyenesre illeszkednek – ez nem más, mint a lineáris korreláció „képe" párhuzamos koordinátákban. Ezt szemlélteti az 5.6. ábra. Figyeljük meg, hogy a párhuzamos egyenesek egymáshoz képest függőlegesen

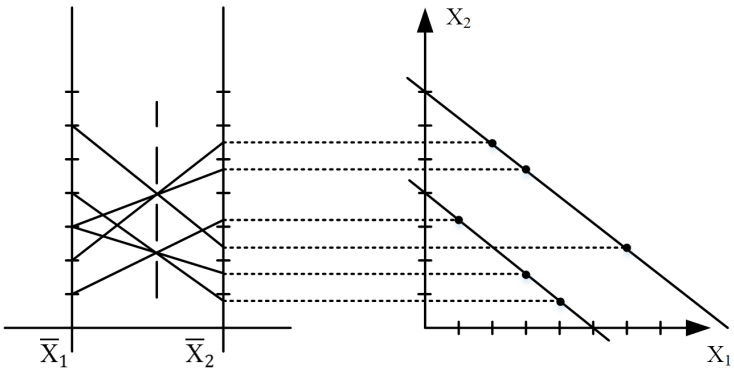
[5] Az ábra forrása: [68], 3. oldal

[6] `iris` néven beépített adatkészlet az R-ben.



5.5. ábra. Az írisz-adatkészlet párhuzamos koordinátákban, fajok szerinti vonaltípusokkal

eltolt pontokként jelennek meg! Könnyű belátni, hogy ehhez hasonlóan egy párhuzamos tengelyek közötti pont vízszintes eltolása pedig a megfelelő egyenes „forgatásának” felel meg.



5.6. ábra. Azonos egyenesre eső pontok leképezése párhuzamos koordinátákra

Az ismert, egyéb alakzatokra vonatkozó, illetve magasabb dimenziószámú minták (mint pl. a hiperbola $\leftrightarrow$ ellipszis leképezés, $\mathbb{R}^3$-ban azonos síkban elhelyezkedés vagy N-dimenziós egyenes – $N-1$ nem párhuzamos hipersík metszete – felismerése) bemutatására itt nincs lehetőségünk; az olvasó figyelmébe ajánljuk a leginkább autoritatívnak tekinthető összefoglaló művet [68].

5.4.3. Eszköztámogatás

Az 5.1. táblázat megadja a bevezetett, illetve hivatkozott diagramtípusokat megvalósító `R` függvényeket. Az `R`-ben különösen a grafika területén igaz, hogy ugyanannak a funkciónak több, egymástól képességekben és kifinomultságban különböző megvalósítása is elérhető. Törekedtünk a legegyszerűbbekre hivatkozásra; mindemellett az olvasó figyelmébe ajánljuk a `ggplot2` [119] és `lattice` [98] csomagokat, melyek hatékony használata bár komolyabb felkészülést igényel, képességeik messze túlszárnyalják az alapvető diagramtípus-megvalósításokat.

5.1. táblázat. Diagramtípusok és R függvények csomagok

Diagramtípus	**függvény**	**csomag**
oszlop, szintvonal, szórás, doboz, hisztogram, mozaik	`barplot, contour, plot, boxplot, hist, mosaicplot`	`graphics`
idősor, hőtérkép	`plot.ts, heatmap`	`stats`
fluktuáció	`fluctile`	`extracat`
párhuzamos koordináták	`parcoord`	`MASS`
szórás-mátrix	`splom`	`lattice`

5.5. Interaktív statisztikai grafika

A számítógéppel megvalósított statisztikai adatvizualizáció lehetőséget teremt arra, hogy egy adatkészlet különböző nézeteivel a felhasználó interakcióba lépjen és ezeknek az interakcióknak *kihatása legyen a többi nézetre*. Az interaktív statisztikai vizualizáció során az eszközök által jellemzően támogatott interakciókat a következőképpen kategorizálhatjuk [110]: *a*) lekérdezések (*queries*); *b*) kiválasztás és csatolt kijelölés (*selection and linked highlighting*); *c*) csatolt elemzések (*linked analyses*) és *d*) helyi interakciók.

A soron következő alfejezetek röviden bemutatják a kategóriákat és a legfontosabb interakciókat. Felhívjuk azonban a figyelmet arra, hogy a különböző eszközök által támogatott interakciókról nem áll módunkban teljes, áttekintő képet adni – az olvasónak javasoljuk a `Mondrian` [109, 110] ingyenes, nyílt forráskódú eszköz megismerését és kipróbálását. Az interaktív statisztikai grafikát ma már több, a vállalati szektornak kínált eszköz is támogatja, az ingyenes és nyílt forráskódúak közül azonban egyértelműen a `Mondrian` a legkiforrottabb. Az `iplots` [115] `R` csomag a `Mondrian`hoz nagyon hasonló képességekkel rendelkezik. Meg kell még említenünk a `GGobit`[24] is, mely a `Mondrian`hoz képest többletfunkciókkal is rendelkezik, kezelése azonban nehézkes és szoftvertechnológiai szempontból is elavultnak tekinthető.

5.5.1. Lekérdezések

A lekérdezések az interaktív diagramon megjelenített elemekkel kapcsolatos statisztikai információk megjelenítését jelentik, jellemzően egér segítségével. A megjeleníthető adatok természetesen az ábra által alkalmazott statisztikai transzformációktól és az aktív kijelölésektől függnek; míg egy szórásdiagramon jellemzően „lekérdezhetjük" pl. egy pont koordinátáit vagy egy kijelölés koordináta-intervallumait, addig egy oszlopdiagramon egy kategória elemszámát és az aktív kijelölésbe tartozó elemek számát. A diagramelem-lekérdezések egy speciális esetét mutatja majd az 5.7. ábra, ahol egy részhalmazra illesztett regressziós egyenes paramétereinek lekérdezése látható.

5.5.2. Helyi interakciók

A diagramokkal önmagukban, a többi diagramra nézve mellékhatásmentesen is interakcióba léphetünk. Egyes operációk – pl. objektumok sorrendjének módosítása, skálamódosítások (ide tartozik a nagyítás is) – általánosan megjelennek; mások ábra-specifikusak. A helyi interakciókkal nem foglalkozunk behatóbban; többségükre tekinthetünk úgy, mint a statikus vizualizáció általában rendelkezésre álló paraméterezési lehetőségeinek „interaktívan" elérhető változataira.

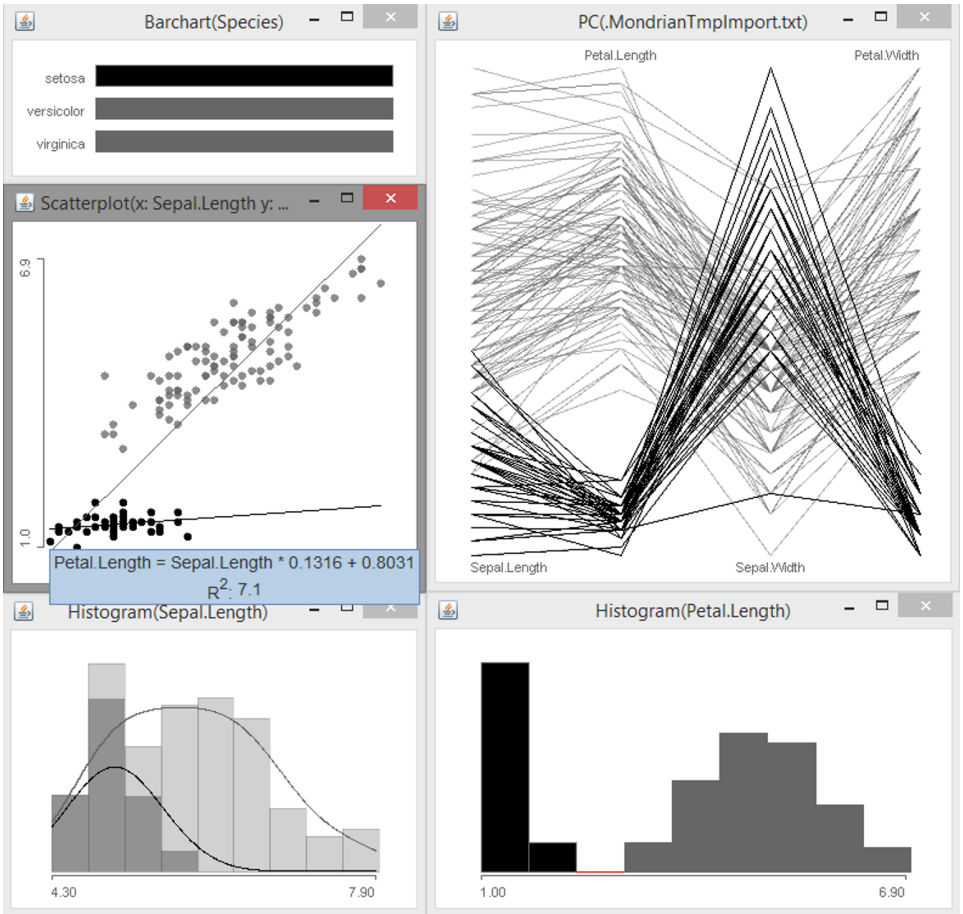
5.5.3. Kiválasztás és csatolt kijelölés

A kiválasztás és csatolt kijelölés az interakció-kategóriák közül a legnagyobb jelentőségű; fogalmazhatunk úgy, hogy ez adja a minőségi különbséget a statikus és az interaktív statisztikai vizualizáció között. Egy adott adatkészleten megjelenített több diagram felfogható úgy, mint egy reláció különböző projekciói (a diagram változóira), majd ezek statisztikai transzformációi. Egy diagramon elemeket (oszlopokat egy oszlopdiagramon, ponthalmazokat egy szórásdiagramon, intervallumot egy dobozdiagramon, ...) jellemzően az egérrel kiválasztva az inverz transzformáció egy „sorhalmazt" határoz meg az eredeti relációban, melynek képe a többi diagramon „kiemelve" vizualizálható.

Erre ad példát az 5.7. ábra a korábban már használt Fisher-féle írisz-adatkészleten, ahol egy, az egérrel „kihúzott" kiválasztó téglalap segítségével a szórásdiagramon választottunk ki pontokat[7]. A kiválasztás motivációja kettős: egyrészt ezek a pontok láthatóan egy elkülönülő klasztert alkotnak a szórásdiagramon, másrészt a kiválasztott csoportra a csészelevél- és sziromhosszúság közötti regressziós kapcsolatot érdemes lenne megvizsgálni. Így ezt a részhalmazt és kapcsolatát a többi megfigyeléssel szeretnénk mélyebben elemezni.

Az adatkészlet egyetlen kategorikus változója, a faj oszlopdiagramján a csatolt kijelölésből rögtön leolvasható, hogy a kiválasztás pontosan az *Iris setosa* faj megfigyelt egyedeit fedi. A párhuzamos koordinátákról leolvasható, hogy ez a faj a sziromhosszúsághoz hasonlóan önmagában a sziromszélességben is szeparálódik a másik kettőtől. Ez a csészelevél

[7] Az eszközökben alapértelmezett esetben a kijelöléseket és a csatolt kiemeléseket karakteres, pl. vörös szín jelöli; a szürke megjelenítés – fekete kiemelés párosra az ábra nyomtatásbaráttá tétele érdekében volt szükség.



5.7. ábra. Kiválasztás és csatolt kijelölés, csatolt elemzés és lekérdezés a `Mondrian`*ban*

jellemzőire nem igaz; felismerhető továbbá, hogy csészelevél-szélesség tekintetében egy megfigyelt egyed a fajon belül kiesőnek (*outlier*) tekinthető. Egyértelmű magas korreláció változópárok között nem olvasható le az adott tengely-permutációnál.

A kiválasztás és csatolt kijelölés egy változatának is tekinthető a kategória-vezérelt színezés, mint interakció (*color brush*). A kategória-vezérelt színezést olyan diagramokról indíthatjuk, melyek kategóriákat jelenítenek meg a vizualizációs primitívként; ilyenek pl. az oszlopdiagram, a mozaik-diagram, de a hisztogram is, ahol az egyes intervallumok oszlopai értelmezhetők kategóriaként. A színezés a kezdeti diagram minden eleméhez – az említett példákon rendre az oszlopokhoz, csempékhez és intervallumokhoz (*bin*-ekhez) – egy színt rendel, melyek segítségével színezi a többi diagram elemeit (pl. egy szórásdiagram pontjait egy-egy színnel, vagy egy oszlopdiagram oszlopait több különböző színű oszlopra bontva, a kategóriák számosságával arányos területtel). Az 5.8. ábra az írisz adatkészlet fajok szerinti színezésére mutat példát.

5.8. ábra. Kategóriánkénti színezés a `Mondrian`*ban*

5.5.4. Csatolt analízis

A kiválasztások hatásának nem feltétlenül kell a (kapcsolt) kiemelésekre korlátozódnia; a kiválasztások változásával reaktívan analízisek, illetve statisztikai modellek felállításai is újrafuthatnak. Ezt az elméleti lehetőséget a jelenlegi eszközök azonban általában nem, vagy csak kevéssé használják ki. A csatolt analízis egyfajta megvalósítása a `Mondrian` szórásdiagramokra illeszthető regressziós egyenes, mely a teljes adatkészlet mellett a mindenkor aktuális kiválasztáson is azonosításra és feltüntetésre kerül. A regressziós modell paraméterei „lekérdezéssel" meg is tekinthetőek (lásd az 5.7. ábra szórásdiagramját).

5.6. Összefoglalás

Az elemzendő adatok „megértésének” és az alapvető összefüggések megsejtésének egyik legfőbb eszköze a statikus és interaktív statisztikai grafika, illetve vizualizáció. A fejezet áttekintést nyújtott ezek legáltalánosabb módszereiről. Nem eshetett szó azonban az adatvizualizáció olyan, jelenleg is aktív kutatás alatt álló területeiről, mint a dinamikus grafika, vagy a szakterületi tudás és statisztikai jellemzők által vezérelt EDA. A további elmélyülést segítendő az érdeklődő olvasó figyelmébe ajánljuk különösen [51] befoglaló könyvét.