Eksploracja Danych

Zagadnienia wizualizacji danych

(c) Marcin Sydow

Zawartość wykładu

- Pryncypia wizualizacji danych
- Wykresy w R
  - pakiet graphics
  - system ggplot2

Pryncypia wizualizacji danych

Wizualizacja graficzna nie ma na celu rozrywki tylko rzetelne przedstawianie danych.

Dobre wizualizacje wymagają kilku czynników, m.in.:

- ciekawa substancja danych
- umiejętności statystyczne
- umiejętności estetyczne

Historyczne przykłady dobrych wizualizacji danych

Edward R. Tufte "The visual display of quantitative information"

- William Playfair
- John Snow "London Cholera Epidemic Map" 1855
- Joseph Minard ("Napoleon Campaign")

Ogólne reguły

- grafika powinna mówić prawdę o danych i przyciągać uwagę do istotnych informacji w tych danych
- tabela całkiem nieźle pokazuje niewielkie zbiory danych (do 20 przypadków)
- wizualna reprezentacja danych powinna być zgodna z reprezentacją numeryczną

(np. postrzegany rozmiar pola koła wydaje się rosnąć wolniej niż faktycznie ($x^{0.8}$))

Zniekształcanie danych przez wizualizację

Miara oszustwa graficznego: (wg Tufte)

$$\textit{miara geometryczna} / \textit{miara numeryczna}$$

Często stosowana zmylająca technika wizualizacji: zmiana dwóch wymiarów reprezentująca zmianę jednego wymiaru w danych

Jeszcze gorzej wygladają reprezentacje 3d jednowymiarowych wielkości (np. baryłki ropy, etc.)

Trzy poziomy

1. ogólna struktura wizualizacji (perspektywa globalna - ogólna)
2. szczegóły wizualizacji (perspektywa lokalna - szczegółowa)
3. dodatkowe, mniej oczywiste informacje wynikające z wizualizacji

Gęstość informacyjna wizualizacji

Oko potrafi rozróżniać bardzo drobne detale (conajmniej 100 na cm^2)

Pewne detale mogą być rozróżnialne gołym okiem nawet do wielkości 0.1mm.

gęstość danych = liczba danych w tabeli/powierzchnia grafiki

Wizualizacja może być zmniejszona, aby powiększyć gęstość.

Do najgęstszych wizualizacji należą dokładne mapy (w tym mapy nieba, etc. nawet do 40K punktów na cm^2)

Jeżeli wizualizacja jest przeładowana szczegółami, to można np. "wygładzać", czy uśredniać, etc.

Jeszcze o estetyce wizualizacji

Elegancja wizualizacji: prostota projektu + złożoność danych

- odpowiedni format
- użycie słów, etykiet objaśniających razem z grafiką
- proporcje, skala
- pewna "historia" do opowiedzenia przez wizualizację
- profesjonalizm elementów i technik
- unikanie niepotrzebnych dekoracji i pustych elementów
- ułatwianie (lokalizacja) ważnych porównań
- tekst i wizualizacje powinny być jednością (nie odseparowane)

Reguły "integralności graficznej"

- graficzna reprezentacja liczb fizycznie zmierzona na powierzchni grafiki (długość, pole powierzchni, etc.) powinna być bezpośrednio proporcjonalna do reprezentowanych wartości liczbowych
- krótkie opisy i objaśniające etykiety tekstowe mogą być dodane do wykresu
- liczba wymiarów graficznych nie powinna przekraczać liczby wymiarów w danych
- kontekst jest kluczowy dla integralności graficznej
- wizualizacja nie powinna być w oderwaniu od kontekstu (np. zmiany wartości w szerszym kontekście czasowym, etc.)
- powinno się ukazywać zmienność w danych nie zmienność w projekcie graficznym
- w wykresach finansowych szeregów czasowych (np. ceny, PKB, etc.) najlepiej używać jednostek standaryzowanych (np. uwzględniających inflację, etc.)

Reguła proporcji "dane-tusz" Tuftego

Reguła:

ilość tuszu związanego z danymi / ilość tuszu wogóle

Należy:

- maksymalizować powyższą miarę.
- używać minimalnie mało "tuszu" do czegokolwiek poza pokazywaniem danych. (ramki, siatki, ozdobniki, etc. - jak najmniej, o ile nie są niezbędne) - usuwać zbędne elementy

(Uwaga: w niektórych sytuacjach powtarzalność jest pozytywna: np. mapy świata, graficzne rozkłady jazdy, etc.)

Ogólne reguły

- przede wszystkim pokazywać dane
- maksymalizować współczynnik dane/tusz
- minimalizować elementy nie związane z danymi
- usuwać powtarzalność (redundantność)
- maksymalizować gęstość danych (na wizualizacji)
- iteracyjnie edytować i poprawiać grafikę

Ornamenty, etc.

Ozdabianie przychodzi łatwiej niż przygotowanie ciekawych danych i ich rzetelnych wizualizacji

Unikać:

- efektu "moire" (interferujące prążki lub kratki, etc.)
- skomplikowanych i absorbujących uwagę deseni.

Kolory

Oko ludzkie nie ma naturalnej zdolności do porządkowania kolorów (choć pewnym pomysłem jest kolejność kolorów tęczy)

Czerwony rzuca się w oczy, ale z kolei daltoniści nie rozróżniają go z zielonym (5-10% populacji)

Niebieski jest odróżniany przez wszystkich.

Czytelność

Miarą nieczytelności wizualizacji jest stopień konieczności dopowiadania.

Bardziej urozmaicone czcionki ułatwiają czytanie (w przec. np. do bezszeryfowej, etc.).

Linie powinny być cienkie.

Grubość pewnych fragmentów linii może podkreślać ważne elementy.

Kształt wizualizacji

Jeśli dane naturalnie sugerują kształt - należy go użyć.

W ogólności raczej pozioma niż pionowa.

- Naturalna zdolność do obserwowania horyzontu i zmian na nim.
- Łatwiej czytać tekst na wizualizacji

Złoty podział

$$\frac{a}{b} = \frac{b}{a+b}$$

1 / 1.618

Inne stosowane podziały: 1/1, 1/1.414, 1/1.732, 1/2

R: dwa pakiety graficzne

Do tworzenia wykresów w R używać można rozmaitych systemów, np:

- pakiet "graphics" (najbardziej standardowy)
- systemy oparte na pakiecie "grid"
  - lattice
  - ggplot2

Funkcja `plot`

Czyści okno i rysuje wykres. Ma wiele argumentów, m.in.[1]:

- x, y (współrzędne x i y, np. jako wektory)
- type (typ wykresu: p,l,b,c, ...)
- main, sub (tytuł i podtytuł)
- xlab, ylab
- asp (ang. aspect ratio)

[1] Funckcja ta jest przeciążona, więc różnie zachowuje się dla różnych obiektów

Inne elementy do uzupełniania wykresu

- legend
- axis
- abline (a,b)
- text (x,y,text)

Rodzaje wykresów w pakiecie graphics

- stripchart(graphics) (wykres paskowy)
- sunflowerplot(graphics) (wykres słonecznikowy)
- persp(graphics) (powierzchnia 3d pod dowolnym kątem)
- contour(graphics) (wykres konturowy)
- pie(graphics) (wykres kołowy)
- barplot(graphics) (wykres słupkowy)
- dotchart(graphics) (wykres kropkowy)
- pairs(graphics) (rozrzuty par)
- coplot(graphics) (rozrzut warunkowy)
- fourfoldplot(graphics) (zależność między 2 zmiennymi)
- contour(graphics) (wykres konturowy)
- filledcontour(graphics) (wypełniony wykres konturowy)
- scatterplot3d(scatterplot3d) (scatter3d(Rcmdr*))
- stars(graphics) (wykres radarowy)

Rysowanie wielu wykresów

- matplot(graphics)
- matlines(graphics)
- matpoints(graphics)

Kolory

Istnieje wiele predefiniowanych zestawów kolorów, np:

- rainbow
- heat.colors
- terrain.colors
- topo.colors
- cm.colors

Listę dostępnych kolorów można otrzymać np. z funkcji "colors()"

Wykresy w innych pakietach

W nawiasach podano nazwę pakietu, w którym występuje funkcja (* - tylko w nowszych pakietach)

- heatmap(stats) ("mapa ciepła")
- parcoord(MASS) (wykres zmian względnych, ang. parallel coordinates plot)
- kde2d(MASS) (jądrowy estymator gęstości)
- plotcorr(ellipse*) (korelacja jako elipsy)
- data.ellipse(car*) (podobnie j.w)
- distplot(vcd*) (dopasowanie do rozkl. dyskr.)
- chplot(chplot*) (wykres otoczkowy)
- bagplot(aplpack*) (uogólnienie boxplot)
- histbackback(Hmisc*) (histogram podwójny)
- rose.diag(circular*) (róża wiatrów)
- faces(aplpack*) (twarze Chernoffa - do 15 zmiennych)

Grafika interaktywna (pakiet: "iplots")

- ibar
- ibox
- ihist
- imap
- imosaic
- ipcp (interaktywne wykresy zmian)

Właściwości linij

(argument "type")

- p - punkt
- l - linia
- c - linia przerywana
- o - linie i punkty
- b - części linij i punkty
- h - pionowe słupki
- S i s - schodki
- n - nic

Wzory linij

(argument "lty")

1. "solid" ciągła
2. "dashed" przerywana
3. "dotted" kropkowana
4. "dotdash" naprzemienna
5. "longdash" inny wariant
6. "twodash" podwójnie przerywana

Ponadto: 0 "blank" (niewidoczna)

Właściwości punktów

Argument "pch"

- wartości 0-25 odpowiadają pewnym znakom graficznym, np.:
  - 19,20 - kropki
  - 21 - okrąg
  - 22,23 - kwadraty
  - 24,25, - trójkąty
- wartośći powyżej 32 odpowiadają kodom ASCII

Niskopoziomowe funkcje graficzne (pakiet "graphics")

Poniższe funkcje pozwalają dorysowywać na wykresie rozmaite elementy, np:

- symbols (rysuje symbole na wykresie)
- polygon (łamane zamknięte)
- segments (odcinki)
- arrows (strzałki)
- curve (krzywa)
- lines (linie łamane)
- rect (prostokąty)
- points (punkty)
- rug (zaznaczanie obserwacji na strzałkach)

Interakcje z wykresami

- identify (identyfikacja na wykresie za pomocą myszki)
- locator (zwracanie współrzędnych z wykresu)

Umieszczanie wielu wykresów na jednym rysunku

Służą do tego instrukcje:

- par(mfrow= c(w,k)) (najpierw wierszami)
- par(mfcol= c(w,k)) (najpierw kolumnami)

Następnie wykonuje się wykresy, np. funkcją "plot", etc.

Bardziej elastyczną funkcją jest "layout".

Parametry graficzne

Można je ustawiać funkcją "par".

Funkcja ta ma kilkadziesiąt możliwych parametrów, które pozwalają szczegółowo kontrolować zachowanie okna graficznego, elementów wykresów, etc.

System ggplot2 (pakiet: ggplot2)

- uproszczona funkcja `qplot` (ang. quick plot)
- rozbudowana funkcja `ggplot2`

Argumenty qplot

- x,y,z (zmienne, które mają być mapowane na 3 osie wykresu)
- data (ramka danych)
- facets (aspekty, do warunkowania)
- margins (czy dodawać rozkłady brzegowe)
- geom (geometria: jak wyświetlać dane)
- stat (jakie podsumowania mają być wyświetlane)

Mapowanie zmiennych

Można mapować do 6 zmiennych, na następujące właściwości:

- x,y,z (współrzędne)
- color (kolor)
- shape (kształt)
- size (wielkość)

Zamapowanie na jakąkolwiek inną właściwość niż współrzędna automatycznie wyświetla legendę.

Wartości właściwości mogą być dobierane automatycznie, aby ułatwić czytelność wykresu.

Geometrie

Geometria decyduje o tym, jak wizualizowane są zmienne.

- jedna zmienna: histogram, freqpoly, density, bar
- dwie zmienne: point, smooth, boxplot, path

Geometrie można parametryzować

Geometrie

geom_NAZWA, gdzie NAZWA to, np.:

- abline, hline, vline
- bar
- contour
- density
- errorbar, errorbarh
- histogram
- linerange
- point, jitter
- rect
- rug
- smooth
- text
- bin2d
- boxplot
- line
- path

Warstwy

Wykresy w systemie ggplot2 składają się z nakładanych kolejno warstw (znakiem "+").

Kolejna warstwa może przesłonić poprzednie warstwy (ale można ustawić "prześwitywanie")

Każda warstwa jest reprezentowana przez obiekt ggplot2, który zawiera:

- dane
- właściwości
- geometrię
- statystyki

Mechanizm warunkowania

Kontrolowany przez parametr `facets`.

Wartość jest w formie formuły. Lewa strona odpowiada wierszom a prawa kolumnom.

Motywy

Kontrolowane przez obiekty `theme_Nazwa`

Przykłady:

- theme_grey (monitor)
- theme_bw (drukarka)

Układy współrzędnych

Przykłady:

- coord_cartesian() (zwykły układ współrzędnych)
- coord_polar() (biegunowy)
- coord_flip() (zamienione osie)
- coord_equal() (te same zakresy obu osi)
- coord_map() (odwzorowania typowe dla map, np. walcowe, etc.)
- coord_trans() (transformacja funkcyjna osi), np.: asn, esp, identity, log, log10, log2, logit, pow10, probit, recip, reverse, sqrt

Nagrywanie wykresów

Grafikę oprócz wyświetlania można też zapisywać do plików w różnych formatach, poprzez nadanie nazwy, formatu, rozmiaru, etc. obrazkowi za pomocą jednej z funkcji:

- png()
- pdf()
- jpeg()
- bitmap()
- postscript()
- pictex() (latex)
- xfig()

Po wywołaniu jednej z powyższych funkcji, należy wykonać rysunek a następnie instrukcję: `dev.off()`, która zapisuje grafikę do wskazanego uprzednio pliku i zamyka urządzenie graficzne.

Kontrolowanie urządzeń graficznych

Przydatne są jeszcze inne funkcje

- dev.cur() (indeks aktualnego urządzenia graficznego)
- dev.list() (lista otwartych urządzeń graficznych)
- dev.next(), dev.prev() (uaktywnia kolejne/poprzednie urządzenie)
- dev.set() (uaktywnia wskazane urządzenie graficzne)

Można też otworzyć nowe okno graficzne: `X11()`

Podsumowanie

- istnieją pewne ogólne pryncypia wizualizacji danych
- w pakiecie R istnieje kilka niezależnych systemów wizualizacji danych, np.:
  - oparty na pakiecie "graphics" (podstawowy)
  - oparty na systemie "ggplot2" (nowszy, oparty na innych mechanizmach)

Przykładowe pytania/zadania/problemy

- jaka jest główna zasada wizualizacji danych?
- wymień kilka podstawowych reguł tworzenia dobrych wizualizacji danych
- czego należy unikać przy wizualizacji danych?
- na czym polega "złoty podział"?
- wymień kilka rodzajów wykresów i opisz każdy z nich
- opisz krótko model wykresu w pakiecie "graphics"
- opisz krótko model wykresu w pakiecie "ggplot2"

Dziękuję za uwagę.