dr Mariusz Piotrowski

Analizy statystyczne i przestrzenne w R

Podstawy analiz i wizualizacji danych.

4 czerwca 2019
wersja robocza

Spis treści

Listings

1. Informacje wstępne.

Kurs podstawowy R można rozpocząć od zainstalowania pakietu w programie R:

Listing 1.1: Interaktywny kurs R - instalacja i uruchomienie.

```
1 install.packages("swirl")
2 library("swirl")
3 swirl()
```

Dodatkowe kursy można doinstalować później. Znajdują się one na stronie https://github.com/swirldev/swirl_courses.

Zagadnienia bazują na różnorodnych materiałach. Podstawowe strony z analizą przestrzenną to:

— http://spatial-analyst.net/wiki/index.php?title=Software
— http://pakillo.github.io/R-GIS-tutorial/
— http://spatial.ly/r/
— http://oscarperpinan.github.io/spacetime-vis/

oraz https://www.r-bloggers.com/the-guerilla-guide-to-r/

2. Konfiguracja środowiska R.

3. Połączenie z bazą Postgresql.

4. Typy danych.

Dane w R przechowywane mogą być w formie:

wektora wartości - kolekcja wartości (np. 1,2,3)należące do tej samej klasy. Mogą to być:
- [**num**] - wartości liczbowe z wartościami dziesiętnymi
- [**int**] - liczby całkowite - bez wartości dziesiętnych
- [**logi**] - operator logiczny - prawda/fałsz
- [**factor**] - zmienne jakościowe (jeśli są uporządkowane wówczas mają klase [**ordered**])

time.series -ts - wektor + zmienna z informacją o dacie

data.frame -(ramka danych) dane przechowywane w układzie:
- kolumny - zmienne
- wiersze - obserwacje

listy - struktura, która pozwala na zagnieżdżanie w niej innych elementów - np. ramek danych, czy innych list.

4.1. Wektor wartości.

Najprostszy sposób tworzenia wektora wartości to użycie funkcji *concatenate* /połącz/ (c).

Listing 4.1: Tworzenie wektora danych funkcją *concatenate* c().

```
1  x <- c(1,2,3,4)
```

4.2. Odwoływanie się do wartości

Indeksowanie danych, czyli odwoływanie się do określonych wartości z wektora, lub ramki danych - odbywa się przez użycie nawiasu kwadratowego [ ].

Wektor wartości Dla wektora [ ] określa pozycję wartości - np:

Listing 4.2: Wskazanie konkretynych wartości z wektora.

```
1  y <- x[c(2,3)]
```

Mogą być użyte ciągi liczbowe:

Listing 4.3: Wskazanie ciągu wartości z wektora.

```
1  y <- x[2:4]
```

Mogą być używane operatory logiczne:

Listing 4.4: Wskazanie wartości przez użycie operatora logicznego.

```
1 y <- x[x > 3]
```

Ramki danych Dla ramek danych podaje się indeks wiersza, następnie indeks kolumny. W pakiecie R wbudowana jest ramka danych - mtcars. Kolejny przykład odwołuje się do tych danych.

Listing 4.5: Wskazanie wartości z ramki danych - ciąg danych.

```
1 z <- mtcars[,1:3]
```

Efektem będzie użycie wszystkich wierszy i kolumn od 1 do 3. W ramach indeksu można filtrować dane używając operatorów logicznych.

Listing 4.6: Wskazanie konkretnych wartości z ramki danych.

```
1  mtcars[mtcars[, "mpg"] > 21,]
```

Funkcja pokaże tylko wiersze, które spełniają warunek - w kolumnie mpg wartości sa większe od 21. Pokazane są w wyniku wszystkie kolumny.

W przypadku, kiedy chcemy odwołać się do jednej kolumny (i uzyskać zamiast ramki danych - wektor) należy użyć znaku $.[1]

Listing 4.7: Wybór jednej kolumny. Wyodrębnienie wektora wartości z ramki danych.

```
1  mtcars$mpg
```

Dodawanie nowej kolumny w ramce danych. Używając funkcji $ można dodawać nowe kolumny.

Listing 4.8: Dodawanie nowej kolumny.

```
1 mtcars$mpgtest <- mtcars$mpg * 2
```

W efekcie powstanie nowa kolumna mpgtest,której wartości są pomnożonymi razy 2 wartościami z kolumny mpg.

Praca na danych sondażowych - case studies Uczestnictwo w kulturze - Katowice Zbudowanie prostej tabel krzyżowej, zgrupowania zmiennych wg jakichś cech - wymaga odwołania się do indeksu.

Z ramki danych (survey-data) chcę zaprezentować dane o wydarzeniu i poziomie wykształcenia uczestników. [,c(...)] oznacza - wybierz WSZYSTKIE wiersze.

Listing 4.9: Identyfikacja numeru i tworzenie ramki danych z dwóch zmiennych.

```
1 which(colnames(survey.data)=="event")
2 which(colnames(survey.data)=="education_level")
3 x <- survey.data[,c(1,13)]
```

[1] W RStudio po wpisaniu znaku $ pojawią się podpowiedzi z nazwami kolumn.

Do analizy danych sondażowych, czyli wszędzie tam, gdzie głównie analizuje się dane mierzone na skalach jakościowych można użyć pakiet «questionr». Przydatne są szczególnie funkcje do analiz procentów i pokazywania procentów w tabelach krzyżowych.

Listing 4.10: Liczenie procentów ze zmiennej.

```
1 library(questionr)
2 freq(survey.data$education_level)
```

Listing 4.11: Liczenie procentów w tabeli krzyżowej.

```
1 library(questionr)
2 x <- survey.data[,c(1,13)]
3 #zrobienie ramki danych z dwoma zmiennymi
4 y <- table(x)
5 #zrobienie tabeli
6 rprop(y)
7 #tabela żkrzyowa z procentami dla wierszy
8 cprop(y)
9 #tabelea żkrzyowa z procentami dla kolumn
```

5. Podstawowe operacje statystyczne.

5.1. Opis jednej zmiennej

Zebrane dane z badań porządkowane są w szeregi statystyczne, dzięki czemu możliwe jest określenie rozkładu wartości, które przyjmuje każda z badanych zmiennych. W tym celu stosuje się różnego rodzaju miary. W celu określenia tego co jest typowe dla zmiennej stosuje się miary skupienia, a w celu ustania wewnętrznej różnorodności zmiennej stosuje się miary rozproszenia.

Statystyki opisowe w R - W przypadku zmienny ilościowych [numeric] podstawowe statystyki można uzyskać funkcją - *summary()*.

Odpowiednikiem jest tzw. pięć liczb Turkeya - *fivenum()*:
— wartość minimalna
— granica pierwszego kwartyla
— mediana
— średnia (tylko w funkcji *summary()*)
— granica trzeciego kwartyla
— wartość maksymalna
W przypadku zmiennych jakościowych funkcja *summary()* pokazuje tablicę częstości zmiennej (uwzględniane są braki danych). Dzięki funkcji *table()* można zbudować tabele kontyngencji (bez braków danych).

Listing 5.1: Podsumowanie statystyk opisowych

```
1 summary(mtcars$mpg)
```

Miary tendencji centralnej

Wielkości średnie Najpopularniejszą miarą skupienia jest średnia arytmetyczna. Mierzy się go jako iloraz sumy wartości pomiarów przez ich liczbę

$$\bar{X} = \frac{\sum x_i}{N}$$

.

W przypadku obliczania **średniej ważonej** każdy pomiar jest mnożony przez wagę, następnie wynik dzielony jest przez liczbę pomiarów. W R średnie można obliczyć następująco:

Listing 5.2: Obiczanie średnich

```
1 mean(dane$zmienna)
2 weighted.mean(dane$zmienna, dane$wagi)
3 1/mean(1/a) #compute the harmonic mean
```

W badaniach społecznych może pojawić się potrzeba użycia innego rodzaju średnich.Warto zwrócić uwagę na:

— średnią odciętą (trymowaną) wyliczana jest średnia odejmując 5% dolnych i 5% górnych wyników
— średnia geometryczna (Wskaźniki w HDI są tak obliczane). Stosowana do określenia przeciętnej wielkości jakiejś zmiany zachodzącej w badanym środowisku. Ale zmiana ma charakter względnie regularny. Np. jakieś zwiększenie cechy rośnie w postępie geometrycznym

$$G = \sqrt[n]{x_1 x_2 ... x_n}$$

i wszystkie $x_i > 0$
— średnia harmoniczna (wykorzystywanej przy obliczeniu średniej liczby mieszkańców na km^2) i

$$H =$$

— średnia krocząca (średnia ruchoma)

.

Wartość środkowa Obliczanie mediany w R

Listing 5.3: Obiczanie wartości środkowej

```
1 median(dane$zmienna)
```

Miary rozproszenia - dyspersji

Odchylenie średnie, przeciętne Miara praktycznie już nie stosowana w statystyce. Odchylenie przeciętne to średnia arytmetyczna wartości bezwzględnych (absolutnych, czyi pomijając znak przed wartością) wszystkich odchyleń poszczególnych wartości pomiarowych od ich średniej arytmetycznej. Uproszczony wzór to

$$d = \frac{\sum |x_i - \bar{x}|}{n}$$

W R do obliczenia można użyć formuły.

Listing 5.4: Obliczanie odchylenia przeciętnego

```
1 mean(abs(dane$zmienna-mean(dane$zmienna)))
```

Wariancja i odchylenie standardowe Wariancja jest obliczania podobnie jak odchylenie przeciętne, jednak zamiast wartości bezwzględnej, natomiast mianownik to liczba obserwacji pomniejszony o 1. (w przypadku obliczeń dla próby) Wzór to

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Odchylenie standardowe jest pierwiastkiem kwadratowym z wariancji. Wzór to

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

W R oblicza się ją jako

Listing 5.5: obliczanie wariancji i odchylenia standardowego

```
var(dane$zmienna)
sd(dane$zmienna)
```

Wadą odchylenia standardowego jest silna podatność na wartości skrajne danej zmiennej.

Współczynnik zmienności - Coefficent of variation Budzącym kontrowersje parametrem określającym rozproszenie jest współczynnik zmienności Pearsona. Zaletą jego jest łatwe obliczenie, oraz zastosowanie do porównań pomiędzy grupami. Do wyliczenia stosuje się iloraz odchylenia standardowego i średniej arytmetycznej

$$V = \frac{s}{\bar{x}}$$

i $\bar{x} \neg 0$. Wartość współczynnika można podać w procentach

$$V_{zmiennej} = \frac{s}{\bar{x}} * 100$$

= % . O kontrowersjach w użyciu tego współczynnika można poczytać u Sørensen.[1]. Aby obliczyć ten współczynnik w R należy zastosować taką formułę:

Listing 5.6: Obliczanie współczynnika zmienności

```
sd(Data$Variable, na.rm=TRUE)/
        mean(Data$Variable, na.rm=TRUE)*100
```

5.2. Opis związku dwóch zmiennych. Zmienne nominalne.

Jakość ustalonej skali rozstrzyga o tym, jakie operacje badawcze moją być wykonane. Jeśli obie zmienne wyrażone są na skalach ilościowych, lub porządkowych wówczas mowa o korelacji, jeśli wyrażone są na skalach nominalnych (jakościowych) mowa o zbieżności, czy asocjacji.

Co ważne - wielkość związku zmiennych w próbie, jedynie określa nam, czy istnieje, czy też nie związek między zmiennymi, nie mówi nam czy przekłada się on na populacje, nie przekłada się on na statystyczną istotność. Test χ^2 określa właśnie to, czy można przenieść wnioski na populację - czy związek jest statystycznie istotny, czy też nie.

[1] https://web.stanford.edu/~sorensen/nomorecv%20revision%20final.pdf

	ilościowa	porządkowa	nominalna
ilościowa	Pearsona współczynnik korelacji (r)		
porządkowa		— Spearmana współczynnik korelacji rangowej/pozycyjnej (R) — τ (tau) Kendalla — Kendalla współczynnik zgodności — γ (gamma Goodmana i Kruskala)	
nominalna	— stosunek korelacyjny η (eta) — współczynnik korelacji dwuseryjnej punktowej		— współczynnik zbieżności/kontyngencji (C), — współczynnik asocjacji (Q) Yula, — współczynnik ϕ (phi) (dla cech zdychotomizowanych) — V Cramera — λ (lambda) Goodmana i Kruskala

5.2.1. Współczynnik asocjacji Q Yule'a

Najprostszym sposobem do obliczeń asocjacji jest współczynnik asocjacji Q - Yule'a (Yule'a - Kendalla), który można stosować wyłącznie do tabel dwudzielnych (2x2), ale w takiej tablicy nie powinny być wartości 0, gdyż współczynnik Q będzie 1, lub -1.

Do obliczeń stosuje się wzór:

$$Q = \frac{ad - bc}{ad - bc}$$

	Wyrzucanie śmieci	Śmiecenie	Suma
Kobiety	18 (a)	7 (b)	25 (a+b)
Mężczyźni	42 (c)	33 (d)	75 (c+d)
Suma	60 (a+c)	40 (b+d)	100 (N)

W efekcie dla tej tabeli dwudzielnej wynik wynosi $Q = 0.33$, czyli jest to związek słaby. Q może przyjmować wartości od -1 do +1. Dodatnie wartości świadczą, że I wariant cechy x współwystępuje z I cechą y, a II wariant cechy x z II wariantem cechy y. Ujemne wartości oznaczają, że I wariant cechy x kojarzy się z II wariantem cechy y, zaś II wariant cechy x, z I wariantem cechy y. ‘

5.2.2. χ^2 (Chi kwadrat) - test niezależności zmiennych

Dla zmiennych jakościowych -kategorialnych lub nominalnych możliwe jest określenie różnicy w rozkładzie zmiennej w badanej próbie. Odbywa się to przez przyjęcie, lub odrzucenie hipotezy zerowej H_0 - w brzmieniu - nie ma statystycznej różnicy w rozkładzie cech zmiennej. χ^2 pozwala na określenie, czy dane w próbie (rozkład według kategorii zmiennej nominalnej) wyniki rozłożyły się wedle proporcji, które są przypadkowe, czy też nie.

Dla małych prób, oprócz wykonania testu na nieciągłość przy obliczaniu χ^2 , należy wykonać test Fishera aby uniknąć błędu z odrzuceniem hipotezy. Test dokładny Fishera przeprowadza się dla bardzo małych prób (np. gdy jedna z liczebności w komórek jest $<$ (mniejsza niż 5) i tablic 2x2. Określa on dokładne prawdopodobieństwo, a nie przybliżone. Test χ^2 może podać prawdopodobieństwo pozwalające na odrzucenie hipotezy zerowej, zaś Test Fishera sprzyja nieodrzuceniu hipotezy zerowej.

Przykładem może być rozkład cechy wyrzucanie śmieci a płeć. Dane prezentuje tabela

	Wyrzucanie śmieci	Śmiecenie	Suma
Kobiety	18 (a)	7 (b)	25 (a+b)
Mężczyźni	42 (c)	33 (d)	75 (c+d)
Suma	60 (a+c)	40 (b+d)	100 (N)

Pierwszym krokiem jest określenie wartości oczekiwanych dla rozkładu cechy. Zasada brzmi: *suma wiersza pomnożona przez sumę kolumny, podzielona przez sumę ogólną.* Można to zrobić wg formuły. Wartość oczekiwana dla komórki a:

$$a_{oczekwiana} = \frac{(a+b)*(a+c)}{N}$$

, czyli

$$a_{oczekiwana} = \frac{25*60}{100} = 15$$

a nowa tabela będzie wyglądała następująco

	Wyrzucanie śmieci	Śmiecenie	Suma
Kobiety	18 (15)	7 (10)	25 (a+b)
Mężczyźni	42 (45)	33 (30)	75 (c+d)
Suma	60 (a+c)	40 (b+d)	100 (N)

Wzór na chi kwadrat prezentuje się następująco:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

gdzie O to wartość obserwowana, a E to wartość oczekiwana.

Dla powyższego przypadku (i prostego zastosowaniu wzoru) $\chi^2 = 2.0$. Po sprawdzeniu w tabeli - dla 1 stopnia swobody[2] - prawdopodobieństwo wynosi 0.16 (więc jest (większe) > od $\alpha = 0{,}05$ lub $\alpha = 0{,}01$) - więc nie odrzucamy H_0. Nie ma istotnej różnicy w sposobie postępowania ze śmieciami ze względu na płeć.

Jeśli p $< \alpha = 0{,}05$ lub $\alpha = 0{,}01$ (czyli jest mniejsze) wówczas hipotezę o niezależności odrzucamy.

W R, aby obliczyć chi kwadrat wystarczy zastosować test.

Listing 5.7: Testowanie chi kwadrat.

```
1 s <-smieci[,c(2,3)]
2 #stworzenie tablicy - indeksowanie
3 chisq.test(s$wyrzucanie.smieci, s$X.1)
4 #przprowadzenie testu
```

Dla powyższych danych R automatycznie dołącza poprawkę Yates'a (dla wartości, które są mniejsze niż 10). Dodatkowo może pojawić się potrzeba przeprowadzenia dokładnego testu Fishera.

Pearson's Chi-squared test with Yates' continuity correction
data: $swyrzucanie.smieciandsX.1$
X-squared = 1.3889, df = 1, p-value = 0.2386

5.2.3. Współczynni phi ϕ

Dla zmiennych uszeregowanych podwójnie(dychotomicznie), kiedy zmienna zawiera tylko dwie klasy - obliczyć można siłę związku obliczając współczynnik ϕ(phi). Problem z tym współczynnikiem polega na tym, że jeśli tabela jest większa to wartość może być większa niż 1, co utrudnia proces interpretacji współczynnika.

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Między ϕi χ^2 istnieje związek bezpośredni

$$\phi^2 = \frac{\chi^2}{N}$$

stąd

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

[2] Stopień swobody oblicza się jako iloczyn liczby kolumn -1 i liczy wierszy -1 (k-1)(w-1)

Sprawia to, że przed określeniem siły związku między zmiennymi w tabeli należy określić, czy jest to związek nie wynikający z przypadku przeprowadzając test na niezależność zmiennych χ^2. Wzór w tej formie może być używany dla tabel większych niż 2x2.

Znak przy ϕnie określa kierunku związku, jak w przypadku korelacji miarowej(Pearsona), jedynie zależy do uporządkowania danych w tabeli 2x2.

5.2.4. Współczynnik kontygencji C - Pearsona

Współczynnik kontygencji w przeciwieństwie do współczynnika ϕ (phi), czy Q Yule'a można stosować dla tabel wielopolowych, bez ograniczenia do wielkości 2x2 (tabeli czteropolowej). Wartości, które przyjmuje współczynnik kontygencji C są z zakresu od 0 do 1.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

5.2.5. V Cramera

Kiedy współczynnik jest wyliczany dla tabeli 2x2 jego wartość jest taka jak we współczynniku kontygencji C. Zastąpił on używany wcześniej współczynnik Czuprowa T

$$V = \sqrt{\frac{\chi^2}{(m-1)N}}$$

m to liczba kolumn, lub wierszy w zależności od tego, która wielkość jest mniejsza

5.2.6. Współczynnik lambda λ Goodmana i Kruskala

Współczynnik bazuje na koncepcji proporcjonalnej redukcji błędu. Zawiera się w przedziale 0 do 1. Miarę tą można obliczać jako:

— symetryczną, wówczas nie ma znaczenia która zmienna jest zależna, a która niezależna

— niesymetryczną, wówczas test jest wykonywany osobno, więc zmiana kolejności będzie skutkowała innymi wynikami

5.3. Opis związku dwóch zmiennych. Zmienne porządkowe.

5.3.1. Współczynnik γ (gamma) Goodmana i Kruskala

W pakiecie SPSS można go używać do wyliczania współczynnika Q Yula.

5.3.2. Współczynnik ρ (rho) Spearmana. Korelacji rang

Współczynnik używany, kiedy dane są porangowane. Porównujemy uszeregowanie dwóch zbiorów danych: obliczamy różnicę rang, podnosimy je do kwadratu, sumujemy. Wartość tego współczynnika zawiera się w przedziale $-1 < 0 < 1$.

5.3.3. Współczynnik τ (tau) b Kendalla. Korelacji rang

Współczynnik używany, kiedy dane porządkowe są mają powiązane rangi.

5.4. Opis związku dwóch zmiennych. Zmienne ilościowe.

5.4.1. Korelacja miarowa/liniowa Pearsona

- Tworzenie macierzy korelacji odbywa się następująco:

Listing 5.8: Korelacja

```
1 cor(mtcars[,1:5])
2 #lub wybór konkretnych kolumn
3 cor(mtcars[,c(1,4)])
```

6. Tworzenie wykresów.

6.1. Pakiet podstawowy.

Prosty mechanizm wizualizacji danych pozwalających na prezentacje statystyk.

6.1.1. Wykres słupkowy

Pozwala na prezentację jednej, lub dwóch zmiennych kategorialnych. Do rysowania wykresu słupkowego służy funkcja *barplot()*.

Częstości sa generowane funkcją *table()*.

Listing 6.1: Rysowanie wykresu słupkowego.

```
1 tab <- table(mtcars$cyl)
2 barplot(tab, horiz = FALSE, las = 1)
```

6.1.2. Wykres pudełkowy.

6.1.3. Wykres kropkowy.

6.2. Pakiet ggplot2.

Tutorial http://r-statistics.co/ggplot2-Tutorial-With-R.html

7. Wizualizacja danych w ggplot2.

Do wizualizacji danych został wykorzystany pakiet ggplot z pakietu R. Podstawowe operacje "czyszczące" i porządkujące dane zostały wykonane w arkuszu kalkulacyjnym, zapisane w formacie csv i zaimportowane do programu RStudio.

7.1. Wykres słupkowy - porównanie danych.

— Rysowanie wykresu w oparciu o przekształcenie danych krótkich "short form" w długie "long form".
— Użycie danych surowych «stat="identity"»
— Obrót wykresu.

```
1 library(reshape)
2 library(ggplot2)
3 dataframe <- melt(krs_osm )
4 dataframe  <- dataframe [complete.cases(krs_osm ),]
5 head(dataframe)
6 ggplot(data = dataframe , aes(x=reorder(Nazwa_indeksu,value), y
     = value, fill = variable), summarise ) + geom_bar(stat="
     identity", width = 0.5) +
7 labs (y="Liczba obiektow", x="Nazwa indeksu infrastruktury",
     size =2, fill='Zrodlo danych' )  +
8 geom_text(aes(label=value), position=position_dodge(width= 0.1 )
     ,hjust = - 0.1, vjust= 0.4,size = 3, color = "black"  ) +
     expand_limits(y=c(0,77000))+
9 coord_flip()
```

7.2. Wykres skrzypcowy i pudełkowy.

```
1 library(ggplot2)
2 ggplot(wydatki_osoba_hist, aes(x=typ_historyczny, y= wydatki)) +
3 geom_violin(aes(fill=typ_historyczny), trim=F) + geom_boxplot(
     width=.2) + xlab("Typ historyczny") + ylab("Wydatki na osobe"
     ) +
4 geom_text(aes(label=gmina),color="blue3", size=3) +
5 geom_rug(sides = 'l')
```

8. Tworzenie map. Pakiet ggplot2.

Kurs online https://www.datacamp.com/courses/working-with-geospatial-data-in-r
Mapa świata https://www.r-bloggers.com/how-to-make-a-global-map-in-r-step-by-ste

Bibliografia

[1] Przemysław Biecek. *Odkrywac! Ujawniac! Objasniac! Zbior esejow o sztuce prezentowania danych*. Fundacja Naukowa SmarterPoland.pl, Warszawa, drugie edition, 2016. ISBN 9788393969500.

[2] Przemysław Biecek. *Przewodnik po pakiecie R*. Oficyna Wydawnicza GiS, Warszawa, 2017.

[3] Roger S Bivand, Edzer J Pebesma, and Virgilio Gómez-Rubio. *Applied spatial data analysis with R*. Springer, New York, Heidelberg, Dordrecht, London, 2013. ISBN 978-1-4614-7618-4. doi: 10.1007/978-1-4614-7618-4. URL `http://link.springer.com/content/pdf/10.1007/978-1-4614-7618-4.pdf`.

[4] Winston Chang. *R Graphics Cookbook*. O'Reilly Media, Inc., 2013. ISBN 1449316956, 9781449316952.

[5] Andrew Field, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications, 2012. ISBN 9781446258460.

[6] Oscar Perpiñán Lamigueiro. *Displaying Time Series, Spatial, and Space-Time Data with R*. Chapman & Hall/CRC, Madrid, 2014. ISBN 9781466565227.

[7] Jesper B Sørensen. The use and misuse of the coefficient of variation in organizational demography research. *Sociological methods & research*, 30 (4):475–491, 2002.

[8] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

[9] Dennis Zielstra and Francesco Tonini. Analysis of Big Spatial Data with PostgreSQL / PostGIS and R – Case Studies in OpenStreetMap and Interactive Web Mapping from R PostgreSQL / PostGIS. In *North Carolina State University Geospatial Analytics Forum*, 2015.