Photo: Vladimir Batagelj, ***Pajčevina***

Multivariatne pajčevine

Vladimir Batagelj

Univerza v Ljubljani

FMF, matematika

Seminar IBMI, Ljubljana, 22. oktober 2004

Sredin seminar IMFM, Ljubljana, 3. november 2004

Kazalo

Multivariatne pajčevine

Naj bo dana *množica večrazsežnih enot* $\mathbf{U}$ in *različnost* d na njej. Zanju lahko določimo dve vrsti omrežij:

Omrežje *k najbližjih sosedov* $\mathbf{G}_N(k) = (\mathbf{U}, A)$

$$(\mathrm{X}, \mathrm{Y}) \in A \Leftrightarrow \mathrm{Y} \text{ je med } k \text{ najbližjimi sosedi enote } \mathrm{X}$$

Usmerjeni povezavi $a(\mathrm{X}, \mathrm{Y}) \in A$ pripišemo utež $w(a) = d(\mathrm{X}, \mathrm{Y})$.
Težave z *enakorazličnimi* enotami – vključimo/izključimo vse; ali določimo dodatno pravilo.

Poseben primer je omrežje *najbližjih sosedov* $\mathbf{G}_N(1)$. Z $\mathbf{G}^*_{NN}$ označimo omrežje, ki vključuje vse enakorazlične enote in z $\mathbf{G}_{NN}$ omrežje, kjer je izbran natanko en sosed.

Omrežje *okolic polmera r* $\mathbf{G}_B(r) = (\mathbf{U}, E)$

$$(\mathrm{X} : \mathrm{Y}) \in E \Leftrightarrow d(\mathrm{X}, \mathrm{Y}) \leq r$$

Najbližjih k sosedov v R-ju

```
k.neighbor2Net <-
# stores network of first  k  neighbors for
# dissimilarity matrix  d  to file  fnet  in Pajek format.
function(fnet,d,k){
  net <- file(fnet,"w")
  n <- nrow(d); rn <- rownames(d)
  cat("*vertices",n,"\n",file=net)
  for (i in 1:n) cat(i," \"",rn[i],"\"\n",sep="",file=net)
  cat("*arcs\n",file=net)
  for (i in 1:n) for (j in order(d[i,])[1:k+1]) {
    cat(i,j,d[i,j],"\n",file=net)
  }
  close(net)
}
```

r-okolice v R-ju

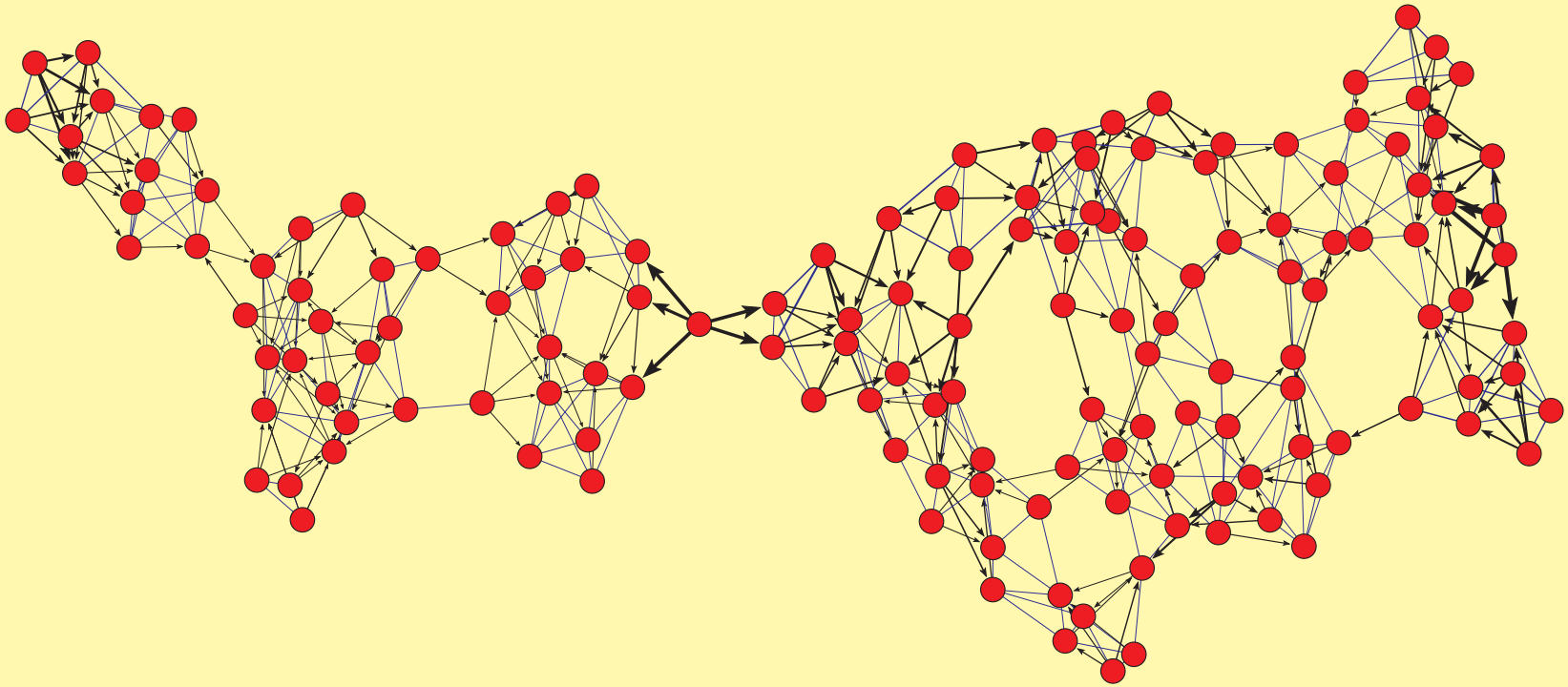
```
r.neighbor2Net <-
# stores network of  r-neighbors (d(v,u) <= r) for
# dissimilarity matrix  d  to file  fnet  in Pajek format.
function(fnet,d,r){
  net <- file(fnet,"w")
  n <- nrow(d); rn <- rownames(d)
  cat("*vertices",n,"\n",file=net)
  for (i in 1:n) cat(i," \"",rn[i],"\"\n",sep="",file=net)
  cat("*edges\n",file=net)
  for (i in 1:n){
    s <- order(d[i,]); j <- 1
    while (d[i,s[j]] <= r) {
      k <- s[j]; if (i < k) cat(i,k,d[i,k],"\n",file=net)
      j <- j+1
    }
  }
  close(net)
}
```

Primer: Fisherjeve perunike

```
stand <-
# standardizes vector  x .
function(x){
  s <- sd(x)
  if (s > 0) (x - mean(x))/s  else  x - x
}

data(iris)
ir <- cbind(stand(iris[,1]),stand(iris[,2]),stand(iris[,3]),
  stand(iris[,4]))
k.neighbor2Net("iris5.net",as.matrix(dist(ir)),5)
```

Omrežje petih najbližjih sosedov



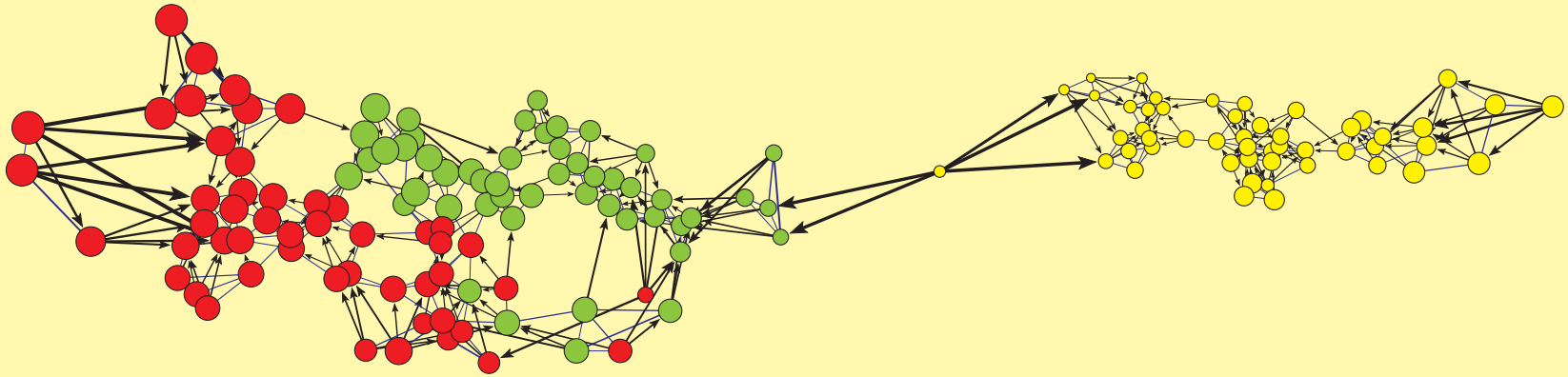
Dopolnitve

```
vector2Clu <-
# stores integer vector  v  as Pajek partition to file  fclu .
function(fclu,v){
  clu <- file(fclu,"w")
  n <- length(v)
  cat("*vertices",n,"\n",file=clu)
  for (i in 1:n) cat(v[i],"\n",file=clu)
  close(clu)
}

vector2Vec <-
# stores vector  v  as Pajek vector to file  fvec .
function(fvec,v){
  vec <- file(fvec,"w")
  n <- length(v)
  cat("*vertices ",n,"\n",file=vec)
  for (i in 1:n) cat(v[i],"\n",file=vec)
  close(vec)
}

vector2Clu("iris.clu",as.numeric(iris[,5]))
vector2Vec("pl.vec",iris[,1])
```

Dopolnjeno omrežje



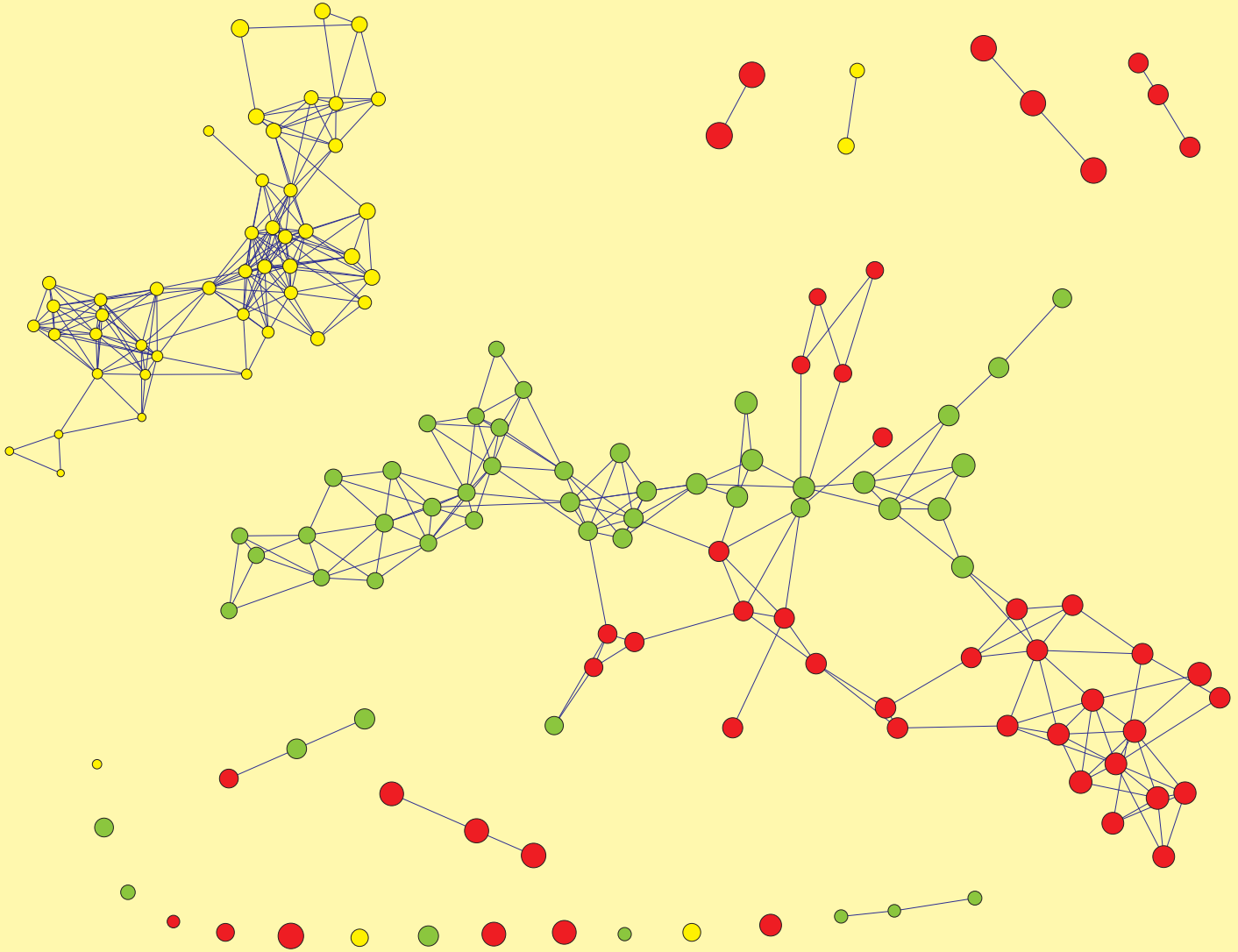
Uteži upoštevane kot različnosti. Velikost točk je sorazmerna $\text{iris}_1 - 4$. Barva točk prikazuje razvrstitev perunik iz podatkov (iris_5).

Okolice

```
r.neighbor2Net("irisR.net",d,0.5)
```

Zgleda, da je omrežje okolic uporabno le v primeru, ko se ’gostota’ posameznih skupin ne spreminja.

Omrežje okolic



Pajčevine in zahtevnost

Problem z večjimi in velikorazsežnimi podatkovji:

Fukunaga, K., Narendra, P.M. (1975), A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, **C-24**, 750-753.

Skiena, Yianilos, Brin, Chua, Chávez &, Murtagh, Dickerson Eppstein, Kleinberg, lectures 22-24, D'haes &, CV Bib: NNC.

Moja predavanja v Konstanzi.

Vmesna rešitev

```
dist2net <-
# Pajek network (in *arcs/*edges format) on file  fnet
# transforms into network with dissimilarities  dist  between
# line endpoints as weights and store it to file  fdis .
function(fnet, fdis, dat, dist, ...){
  net <- file(fnet,"r"); dis <- file(fdis,"w"); copy <- TRUE
  while (copy) {
    t <- readLines(net,n=1)
    cat(t,"\n",file=dis)
    if (length(t)>0){
      if (substr(t,1,1) == "*") {
        c <- tolower(substr(t,2,2))
        copy <- (c != "e") & (c != "a")
        if (copy & (c != "v")) cat("Unexpected command in NET\n")
      }
    } else copy <- FALSE
  }
```

... Vmesna rešitev

```
  copy <- TRUE
  digits <- c("0","1","2","3","4","5","6","7","8","9","-")
  while (copy) {
    t <- readLines(net,n=1)
    if (length(t)>0){
      if (substr(t,1,1) == "*") {
        c <- tolower(substr(t,2,2))
        copy <- (c == "e") | (c == "a")
        if (copy) cat(t,"\n",file=dis)
      } else {
        z <- unlist(strsplit(t," ")); z <- z[z != ""]
        u <- as.numeric(z[1]); v <- as.numeric(z[2])
        d <- dist(rbind(dat[u,],dat[v,]),...)
        j <- 3; if (is.element(substr(z[3],1,1),digits)) j <- 4
        cat(z[1],z[2],d,file=dis)
        if (j <= length(z)) for(i in j:length(z)) cat(z[i],file=dis)
        cat("\n",file=dis)
      }
    } else copy <- FALSE
  }
  close(dis); close(net)
}
```

Primer: Slovenske občine

Primer: Slovenske občine 2002

- V_1 – ime
- V_2 – povprečna velikost gospodinjstva
- V_3 – povprečno stanovanj na stavbo
- V_4 – povprečna neto plača

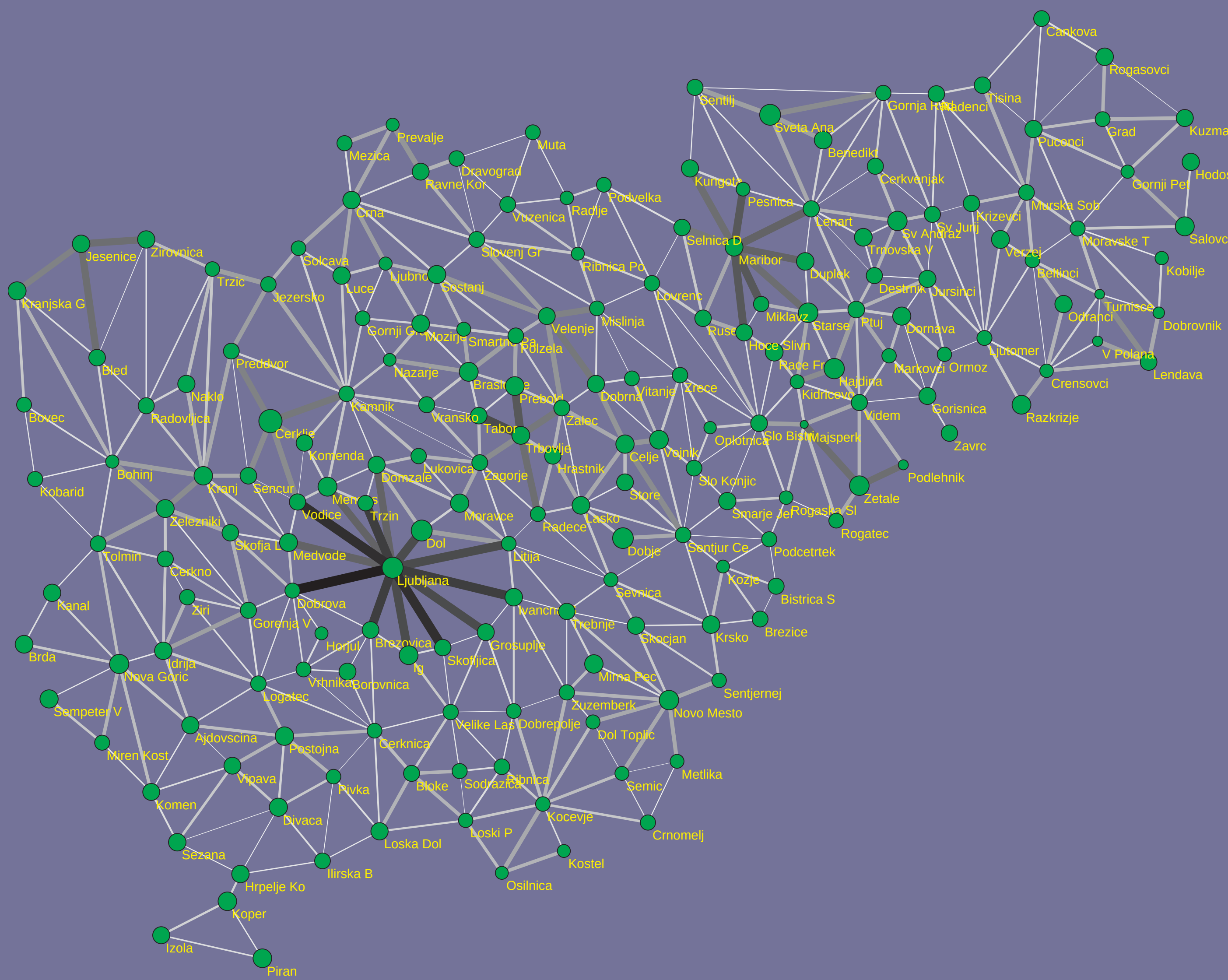
Okolice

```
o <- read.delim("obctest.txt",header=FALSE,sep="\t",dec=",")
or <- cbind(stand(o[,2]),stand(o[,3]),stand(o[,4]))
dist2net("obcxy.net","ObcineDis.net",or,dist)
vector2Vec("placa.vec",o[,4])

dist2net("obcxy.net","ObcineMan.net",or,dist,method="manhattan")
```

Velikost točk je sorazmerna $V_4 - 70000$.

Občine


Cankova
Rogasovci
Tisina
Radenci
Sentilj
Sveta Ana
Grad
Kuzma
Puconci
Benedikt
Hodos
Gornji Pet
Cerkvenjak
Kungota
Pesnica
Murska Sob
Krizevci
Lenart
Sv Andraz
Sv Jurij
Salovci
Selnica D
Moravske T
Trnovska V
Verzej
Maribor
Duplek
Kobilje
Beltinci
Destrnik
Jursinci
Turnisce
Lovrenc
Odranci
Miklavz
Starse
Ptuj
Dornava
Dobrovnik
Hoce Slivn
V Polana
Ljutomer
Markovci
Ormoz
Lendava
Prevalje
Mezica
Muta
Dravograd
Ravne Kor
Podvelka
Radlje
Vuzenica
Crna
Slovenj Gr
Ribnica Po
Jesenice
Zirovnica
Solcava
Trzic
Jezersko
Luce
Sostanj
Kranjska G
Velenje
Mislinja
Polzela
Preddvor
Nazarje
Bled
Vitanje
Zrece
Dobrna
Prebold
Naklo
Kamnik
Vransko
Radovljica
Bovec
Tabor
Zalec
Oplotnica
Majsperk
Komenda
Trbovlje
Vojnik
Celje
Videm
Gorisnica
Razkrizje
Zavrc
Bohinj
Kobarid
Kranj
Sencur
Domzale
Lukovica
Zagorje
Hrastnik
Slo Konjic
Podlehnik
Zetale
Vodice
Trzin
Moravce
Store
Zelezniki
Rogaska Sl
Rogatec
Medvode
Dol
Radece
Lasko
Dobje
Sentjur Ce
Podcetrtek
Tolmin
Cerkno
Litija
Ljubljana
Kozje
Bistrica S
Sevnica
Kanal
Ziri
Dobrova
Trebnje
Skocjan
Krsko
Brezice
Gorenja V
Horjul
Brezovica
Grosuplje
Brda
Idrija
Ig
Skofljica
Nova Goric
Vrhnika
Borovnica
Mirna Pec
Sentjernej
Logatec
Sempeter V
Zuzemberk
Novo Mesto
Velike Las
Dobrepolje
Dol Toplic
Miren Kost
Ajdovscina
Cerknica
Postojna
Vipava
Pivka
Bloke
Sodrazica
Ribnica
Semic
Metlika
Komen
Divaca
Kocevje
Crnomelj
Loski P
Sezana
Loska Dol
Ilirska B
Kostel
Hrpelje Ko
Osilnica
Koper
Izola
Piran

Zaključek

- R-project, R for Windows.
- Pajek.
- NetAna.