A Geostatistical Analysis of Water Losses in an Urban Utility using the R and Tinn-R Open-Source Statistical Computing Environments

F.Kizito[a], G. Ngirane-Katashaya[b] & R. Thunvik[a]

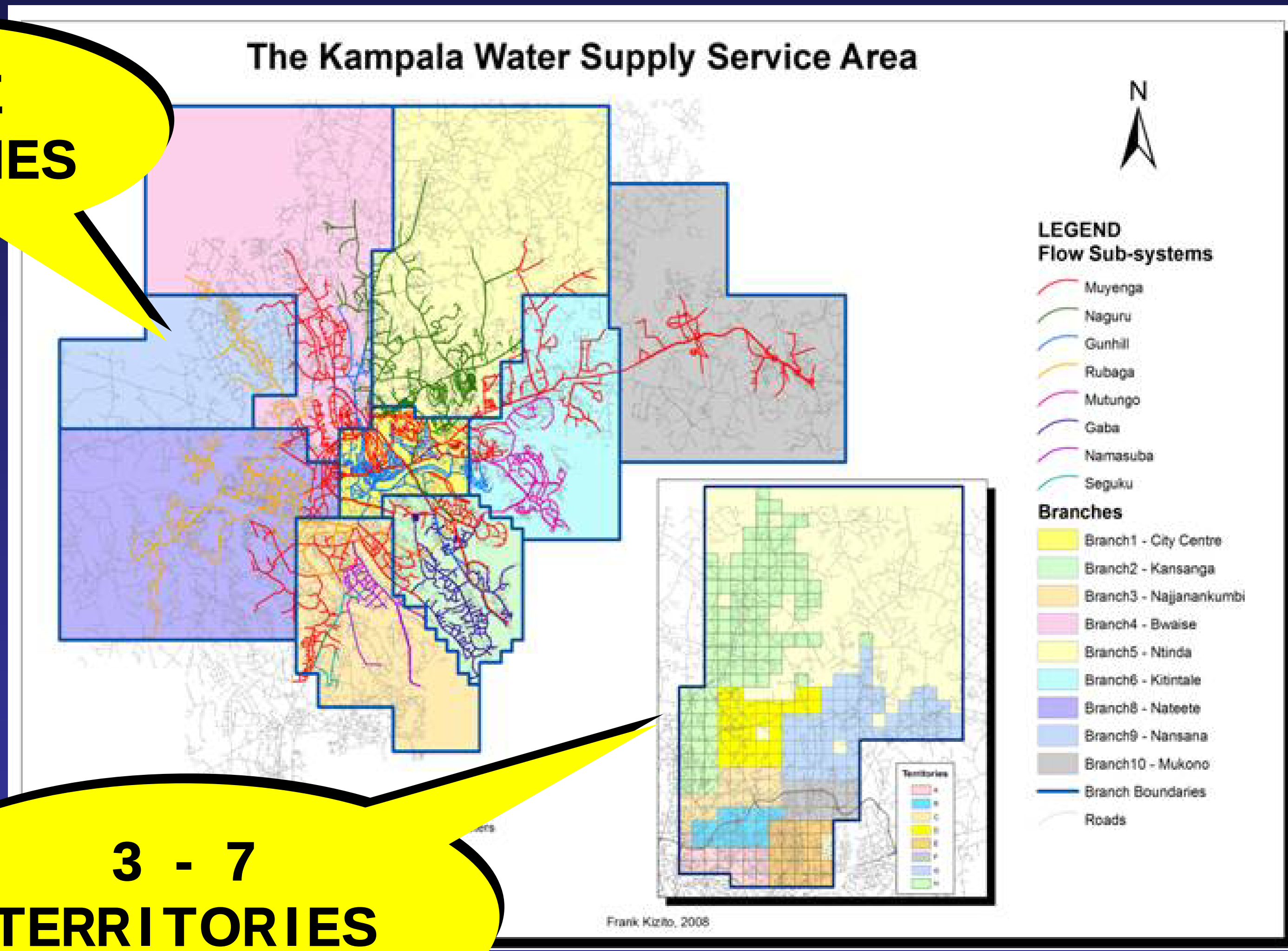
[a]Royal Institute of Technology (KTH), Sweden

[b]Makerere University, Uganda



THE STUDY AREA:

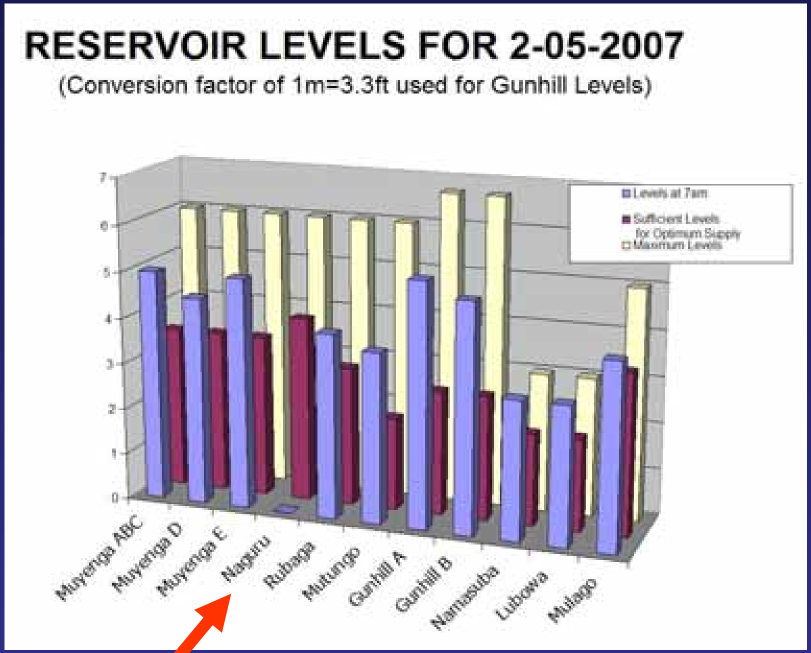
THE STUDY AREA:

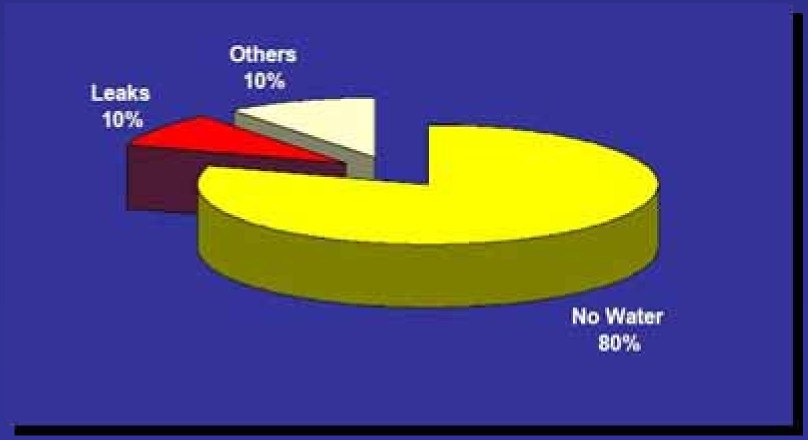


NINE BRANCHES

3 - 7 TERRITORIES PER BRANCH

PROBLEM ISSUES:





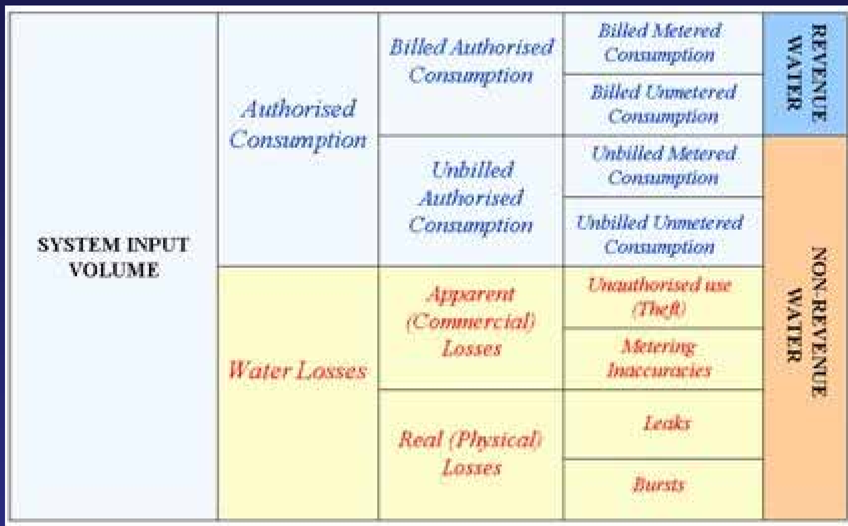
Problem 1:

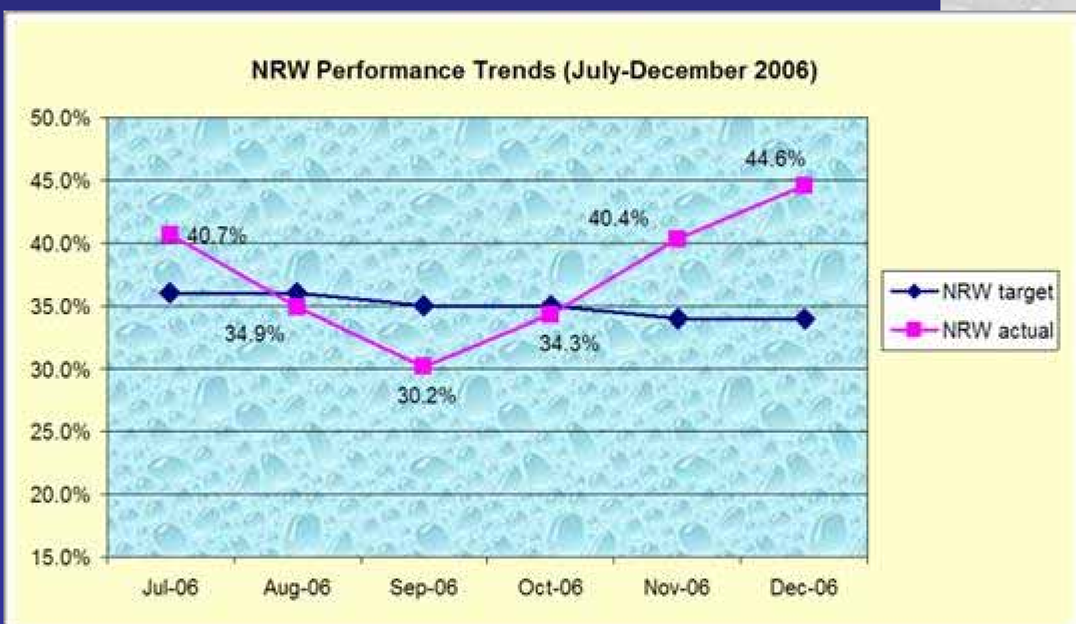
Water Supply Intermittence

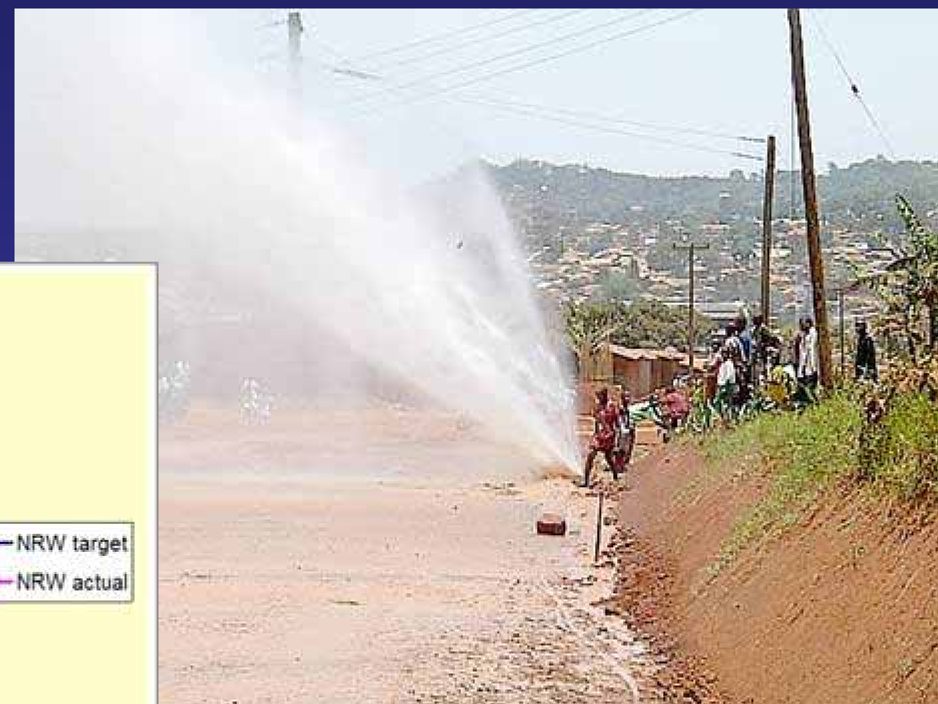
HAPHAZARD PIPE INFRASTRUCTURE
(Photo: KW Archives)

PROBLEM ISSUES:

Problem 2:
High Non-Revenue Water (NRW) levels



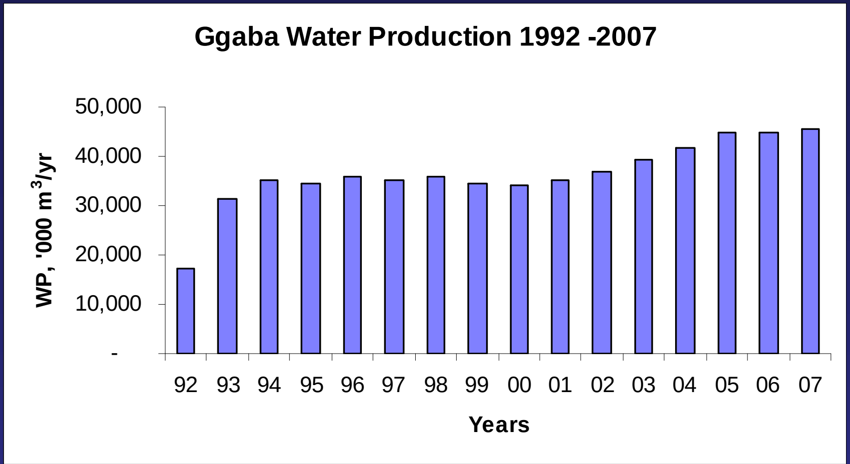


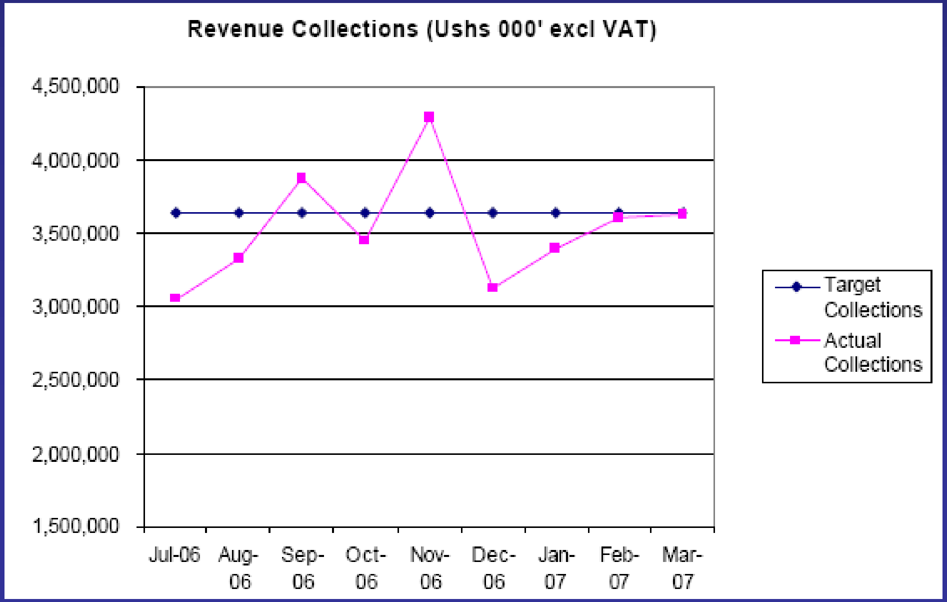


LEAKS AND BURSTS
(Photo: KW Archives)

PROBLEM ISSUES:

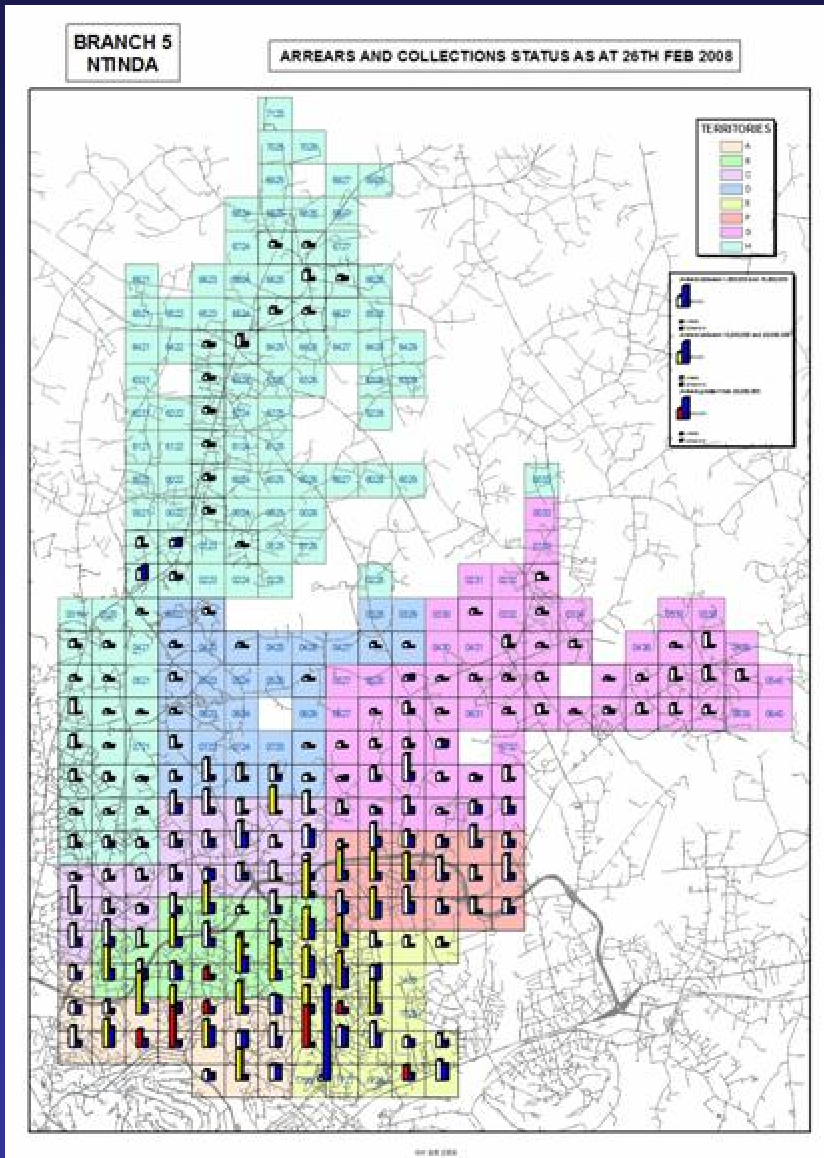
Problem 3:
Rising Production Costs vs Stagnant Collections





GABA III WATER WORKS
(Photo: Roger Thunvik)

PROBLEM ISSUES:

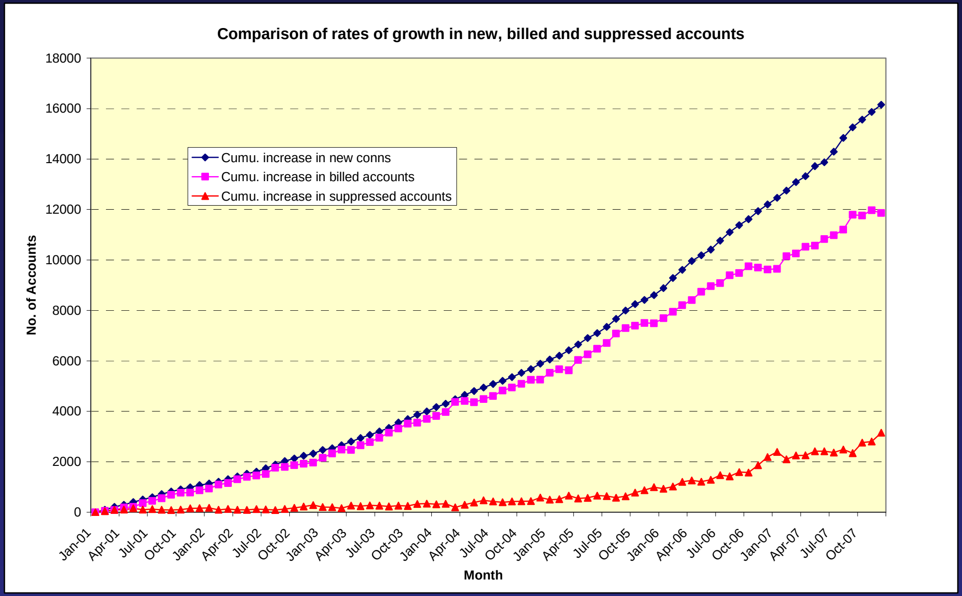


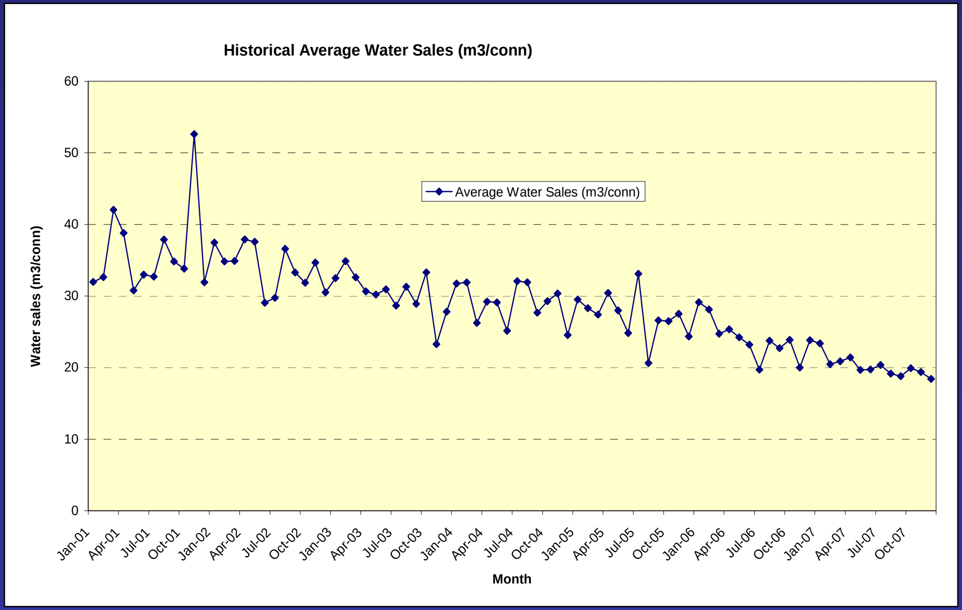
Problem 4:

Increase in disconnections due to non-payment

A TYPICAL BRANCH OFFICE
(Photo: KW Archives)

STUDY OBJECTIVE:





To investigate the possible causes of declining water sales in a selected Branch, using geostatistical analysis.

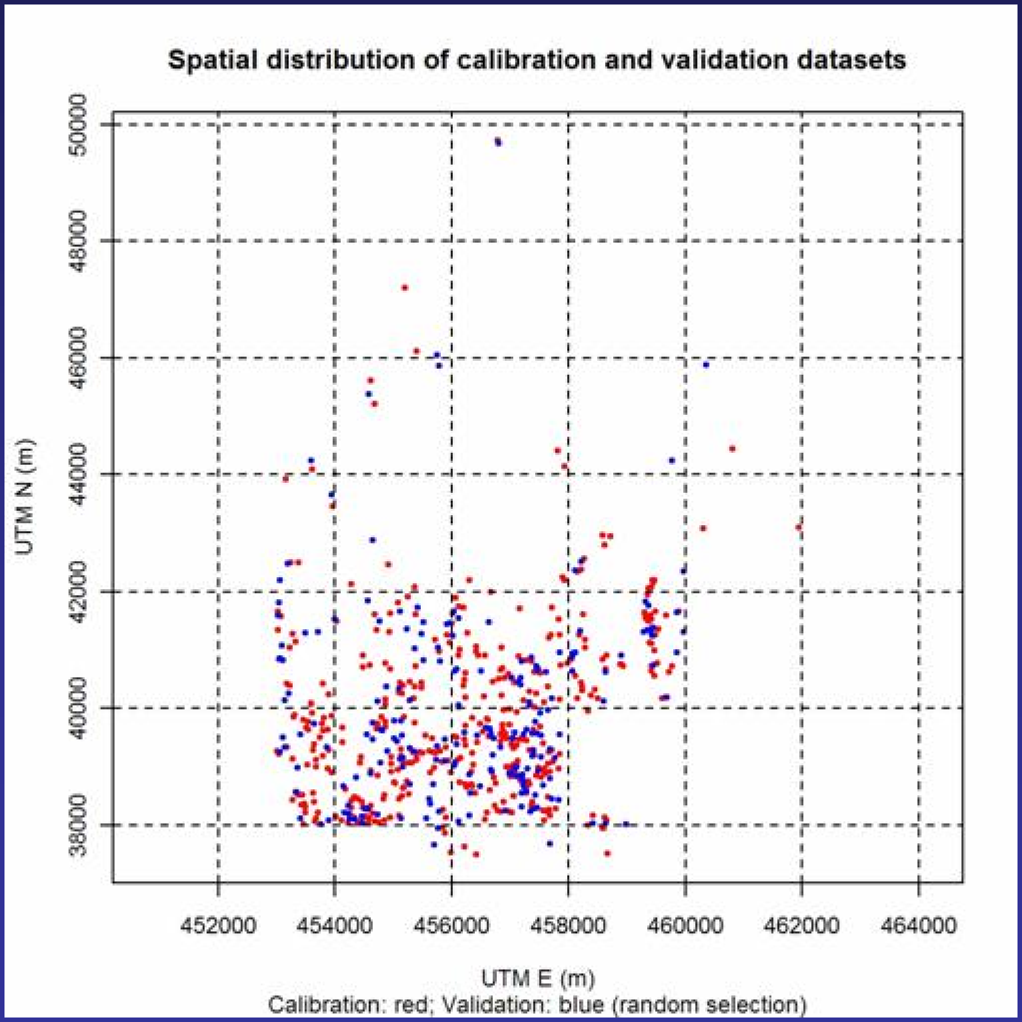
TOOLS USED:

R version 2.6.1 (2007-11-26)
Copyright (C) 2007 The R Foundation for Statistical Computing

R: A statistical computing environment, developed under the Free Software Foundation's GNU project. (http://www.r-project.org/; http://www.gnu.org/)

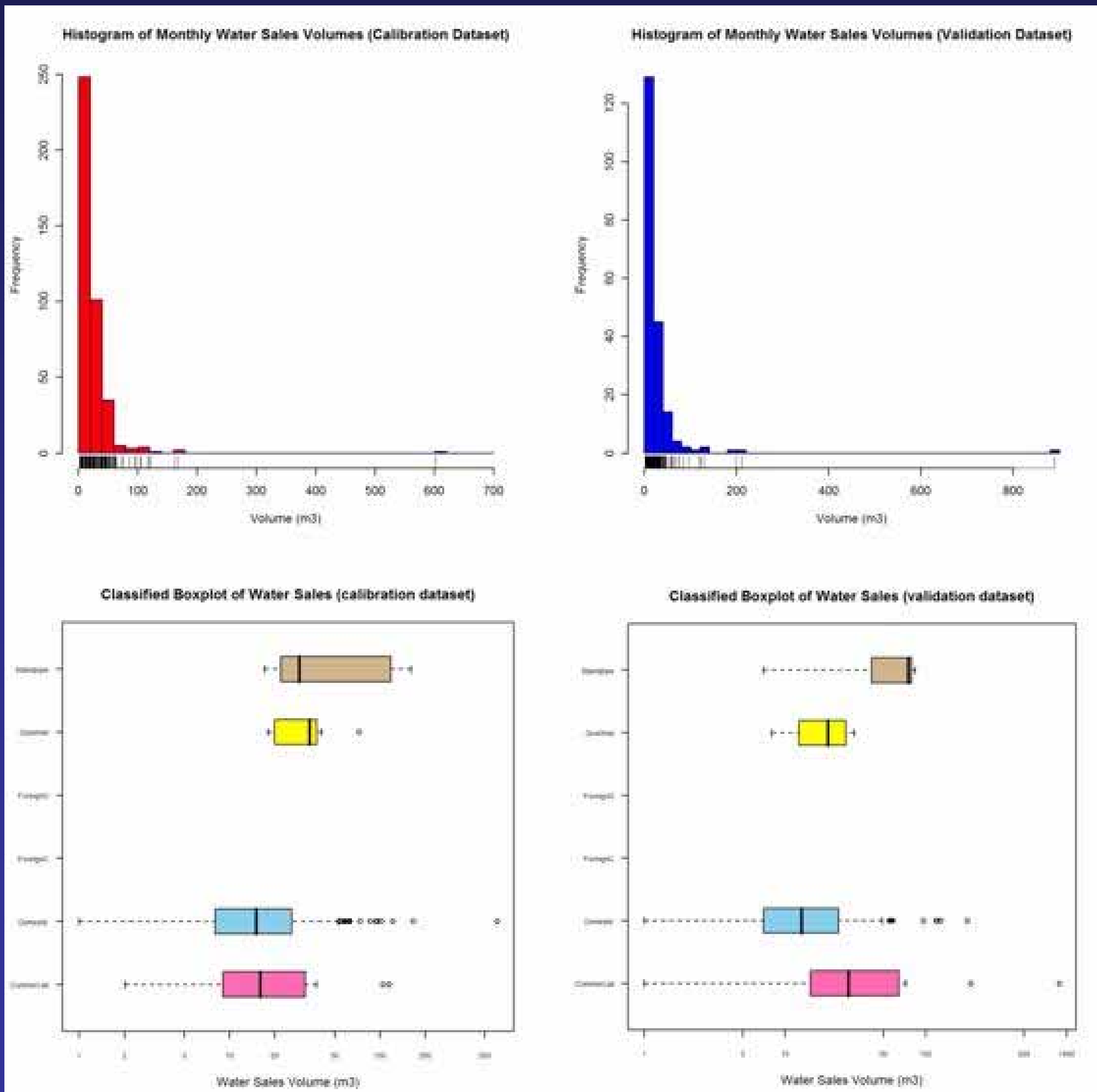
Tinn R: A free, simple code editor for R. (http://sourceforge.net/projects/tinn-r/; http://www.sciviews.org/Tinn-R/)

GEOSTATISTICAL ANALYSES: (1) Selection of calibration and validation datasets



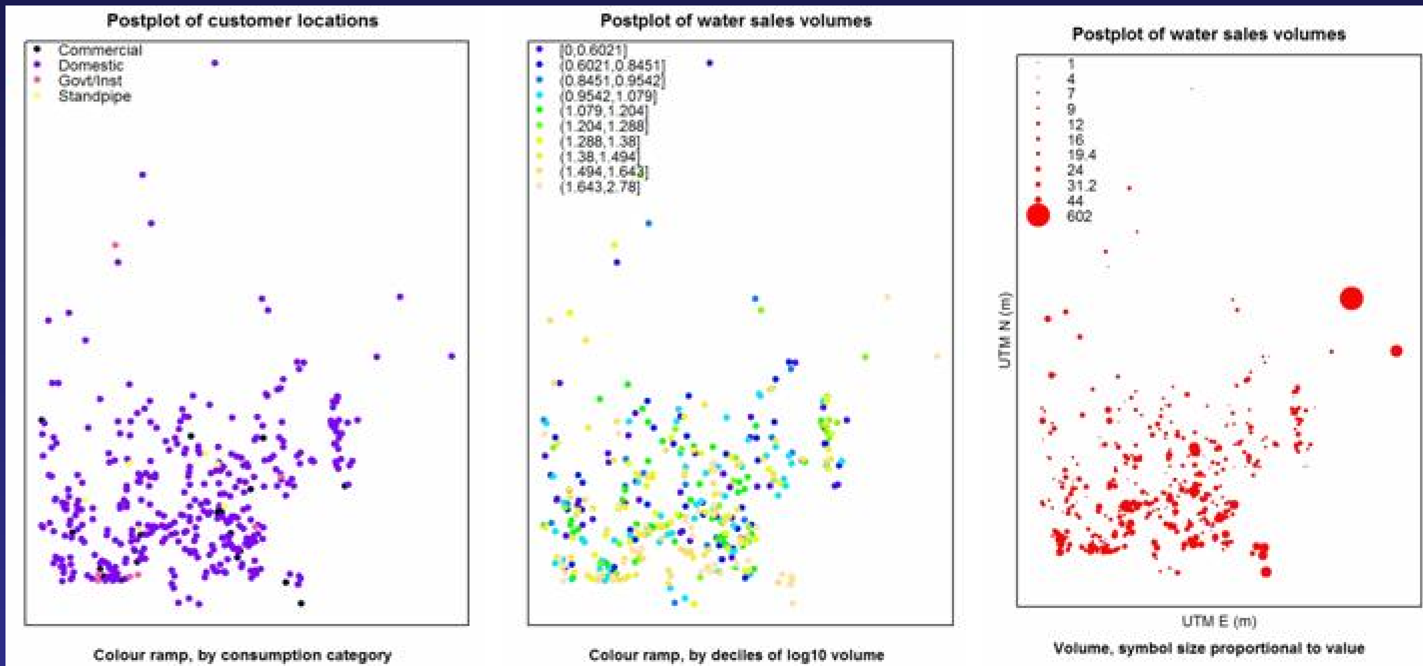

```
# Random selection of calibration and validation datasets

cal.id <- sample(1:dim(d200312.all@data)[1], 4592)

sort(cal.id)

d200312.cal <- d200312.all[cal.id, ]

str(d200312.cal)

summary(d200312.cal)

d200312.val <-
d200312.all[!is.element(1:dim(d200312.all@data)[1], cal.id),
]

str(d200312.val)

summary(d200312.val)

# Save calibration and validation datasets

save(d200312.all, d200312.cal, d200312.val, file =
"Project.RData")

# Spatial distribution of calibration and validation datasets

plot(coordinates(d200312.cal), col = "red", asp = 1, pch =
20,cex = 0.8, xlab = "UTM E (m)", ylab = "UTM N (m)", main
= and validation datasets", sub = selection)")

points(coordinates(d200312.val), col = "blue", pch = 20, cex
= 0.8)

grid(lty = 500, col = "black")
```

GEOSTATISTICAL ANALYSES: (2) Exploratory Data Analysis



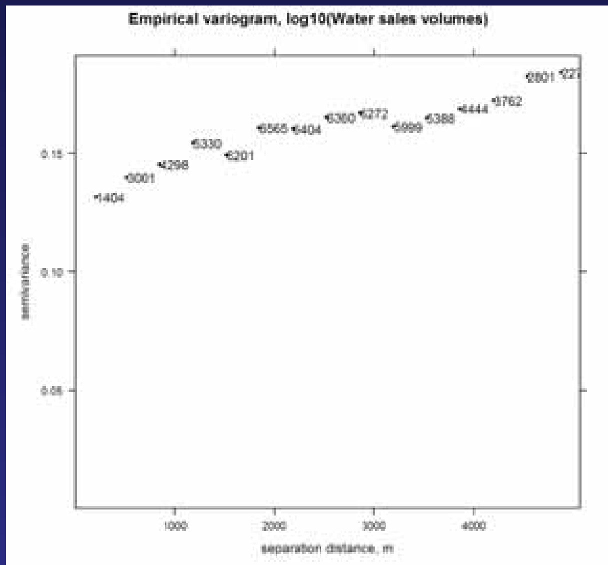

```
# Univariate exploratory graphics

#(a) Histograms

hist(d200312.cal$V200312, breaks = seq(0, 700, by = 20), col = "red",border = "darkblue", main = Dataset)",xlab = "Volume (m3)")

rug(d200312.cal$V200312)

hist(d200312.val$V200312, breaks = seq(0, 900, by = 20), col = "blue",border = "black", main = "Histogram of Monthly Water Sales Volumes (Validation Dataset)", xlab = "Volume (m3)")

rug(d200312.val$V200312)

#(b) Classified Boxplot

par(las=1)

par(cex.axis=0.6)

boxplot(d200312.cal$V200312 ~ d200312.cal$CCat, horizontal = T, outline = T, log="x", pars = list(boxwex = 0.4, staplewex = 0.3, outwex = 0.3), col = c("hotpink", "skyblue", "azure", "turquoise", "yellow","tan"), main = "Classified Boxplot of Water Sales (calibration dataset)", xlab = Volume (m3)",)
```

GEOSTATISTICAL ANALYSES: (3) Investigation of spatial structure

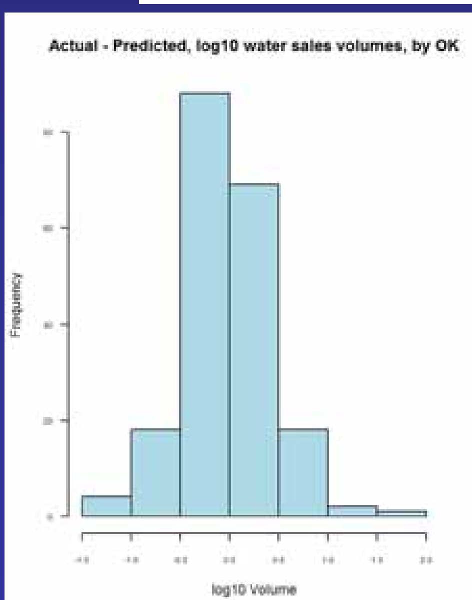


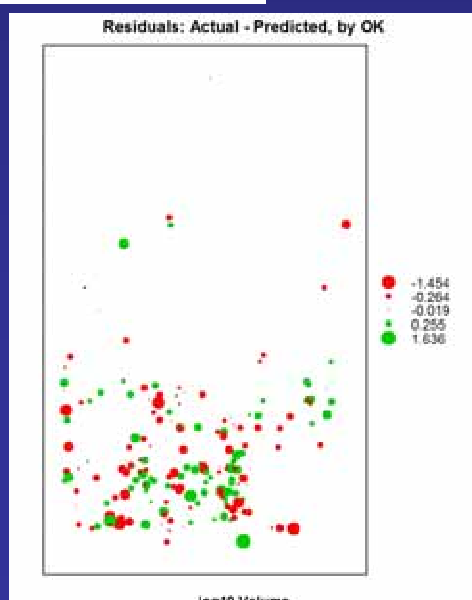

```
# Postplot of consumption categories

print(spplot(d200312.cal, z = "CCat", col.regions = bpy.colors(6), main = "Postplot of customer locations", sub = "Colour ramp, by consumption category", key.space = "inside", cex = 0.8))

# Postplots of water sales values

#(a) Colour ramp, by deciles of log10 volume

print(spplot(d200312.cal, z = "log10V200312", col.regions = topo.colors(64), main = "Postplot of water sales volumes", sub = log10 volume", key.space = "inside", cuts = quantile(d200312.cal$log10V200312, seq(0, 1, by = 0.1)), cex = 0.8))

#(b) Symbol size representing log10 volume

print(bubble(d200312.cal, "V200312", main = "Postplot of water sales volumes", sub = "Volume, symbol size proportional to value", col = "red", scales = list(draw = TRUE), xlab = "UTM E (m)", ylab = "UTM N (m)", fill = T, aspect = "iso", do.sqrt = T, key.entries = quantile(d200312.cal$V200312, seq(0, 1, by = 0.1))))
```

GEOSTATISTICAL ANALYSES: (4) Variogram modelling

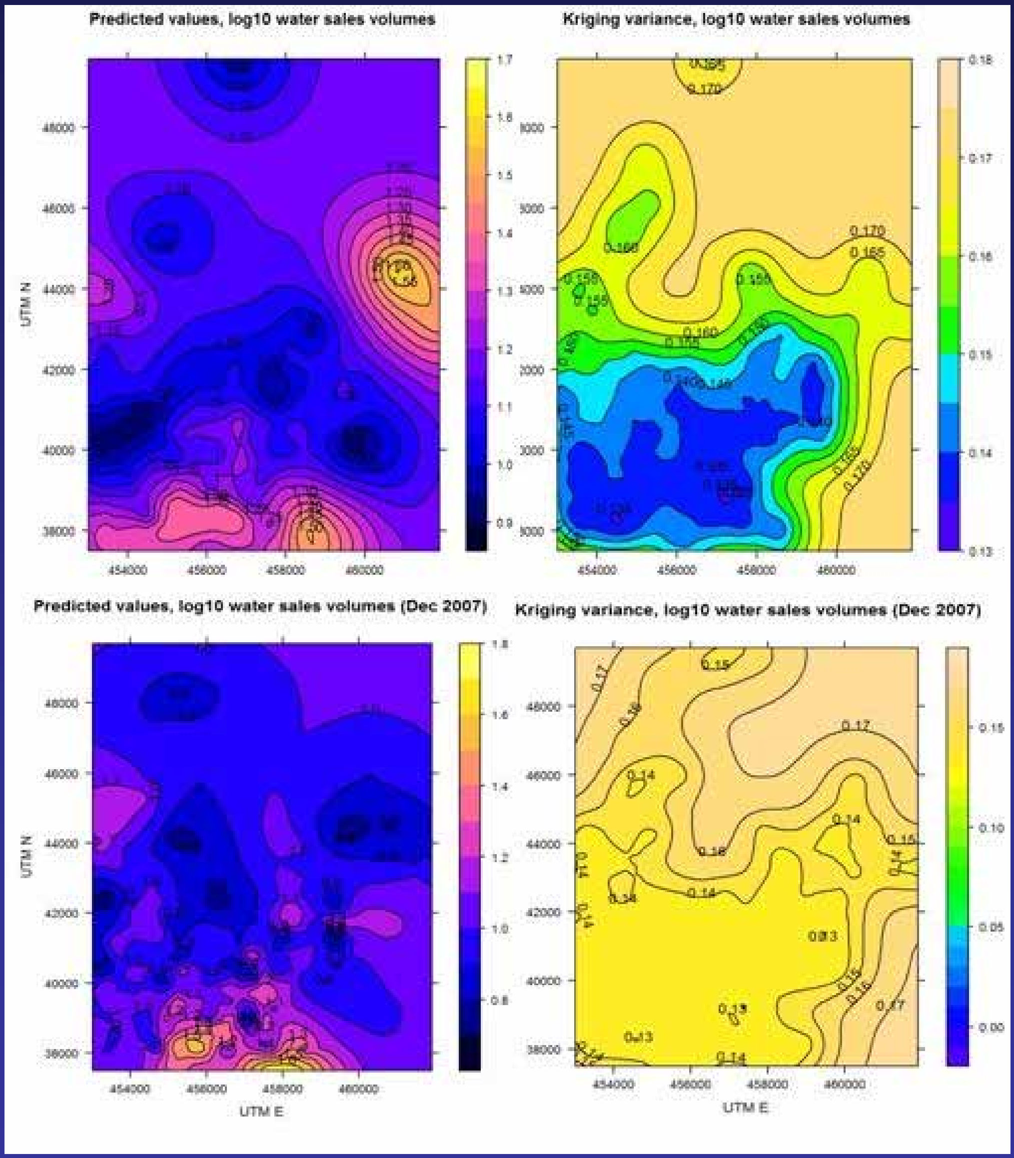



```r
# Empirical variogram of logarithm of water sales volumes
v <- variogram(log10(V200312) ~ 1, loc = d200312.cal)
print (v)
print(plot(v, plot.numbers = T, main = "Empirical variogram, log10(Water sales volumes)", xlab = "separation distance, m", col = "darkblue", pch = 20))
# Fitting variogram model
vm <- vgm(0.04, "Sph", 1200, 0.13)
vmf <- fit.variogram(v, vm)
(vm <- vgm(0.005, "Sph", 5000, add.to = vmf))
(vmf2 <- fit.variogram(v, vm))
# Model validation using validation dataset
#(a) Add log10 volume field to validation dataset
d200312.val$log10V200312 <- log10(d200312.val$V200312)
#(b) Predict log10 water sales volumes at validation points
k.val <- krige(log10(V200312) ~ 1, loc = d200312.cal, newdata = d200312.val, model = vmf2)
summary(k.val)
#(c) Compare predictions to actual values
resid <- d200312.val$log10V200312 - k.val$var1.pred
summary(resid)
hist(resid, main = "Actual - Predicted, log10 water sales volumes, by OK", xlab = "log10 Volume", col = "lightblue")
resid <- as.data.frame(resid)
coordinates(resid) <- coordinates(d200312.val)
bubble(resid, maxsize= 2, main = "Residuals: Actual - Predicted, by OK", sub = "log10 Volume")
```





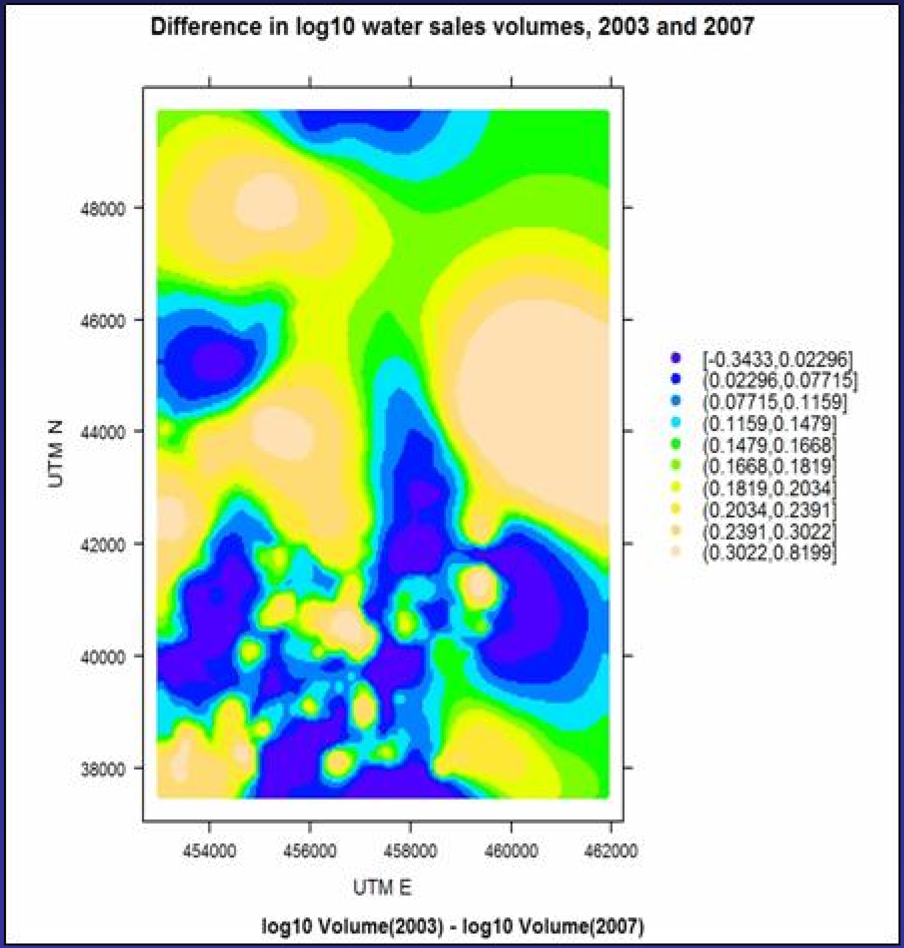
GEOSTATISTICAL ANALYSES: (5) Water sales kriging predictions




```
# Raster map of predictions
#(a) Selection of grid spacing
vc <- variogram(log10(V200312) ~ 1, loc = d200312.cal, cloud = T)
str(vc)
min(vc$dist)
(min.x <- floor(bbox(d200312.cal)["XCoord", "min"] - 10))
(min.y <- floor(bbox(d200312.cal)["YCoord", "min"] - 10))
(max.x <- ceiling(bbox(d200312.cal)["XCoord", "max"] + 10))
(max.y <- ceiling(bbox(d200312.cal)["YCoord", "max"] + 10))
(cells.x <- (max.x - min.x)/50)
(cells.y <- (max.y - min.y)/50)
cells.x * cells.y
#(b) Creation of grid
d200312.grid <- SpatialGrid(GridTopology(c(min.x + 25, min.y + 25), c(50, 50),
c(cells.x, cells.y)))
str(d200312.grid)
d200312.grid <- SpatialPoints(d200312.grid)
gridded(d200312.grid) <- T
str(d200312.grid)
#(c) Prediction on grid
k.grid <- krige(log10(V200312) ~ 1, loc = d200312.cal, newdata = d200312.grid, model = vmf2)
summary(k.grid)
#(d) Plotting gridded predictions
print(spplot(k.grid, zcol = "var1.pred", pretty = T, contour = T, col.regions = bpy.colors(64), main = "Predicted values, log10 water sales volumes", xlab = "UTM E", ylab = "UTM N", scales = list(draw = T)))
```

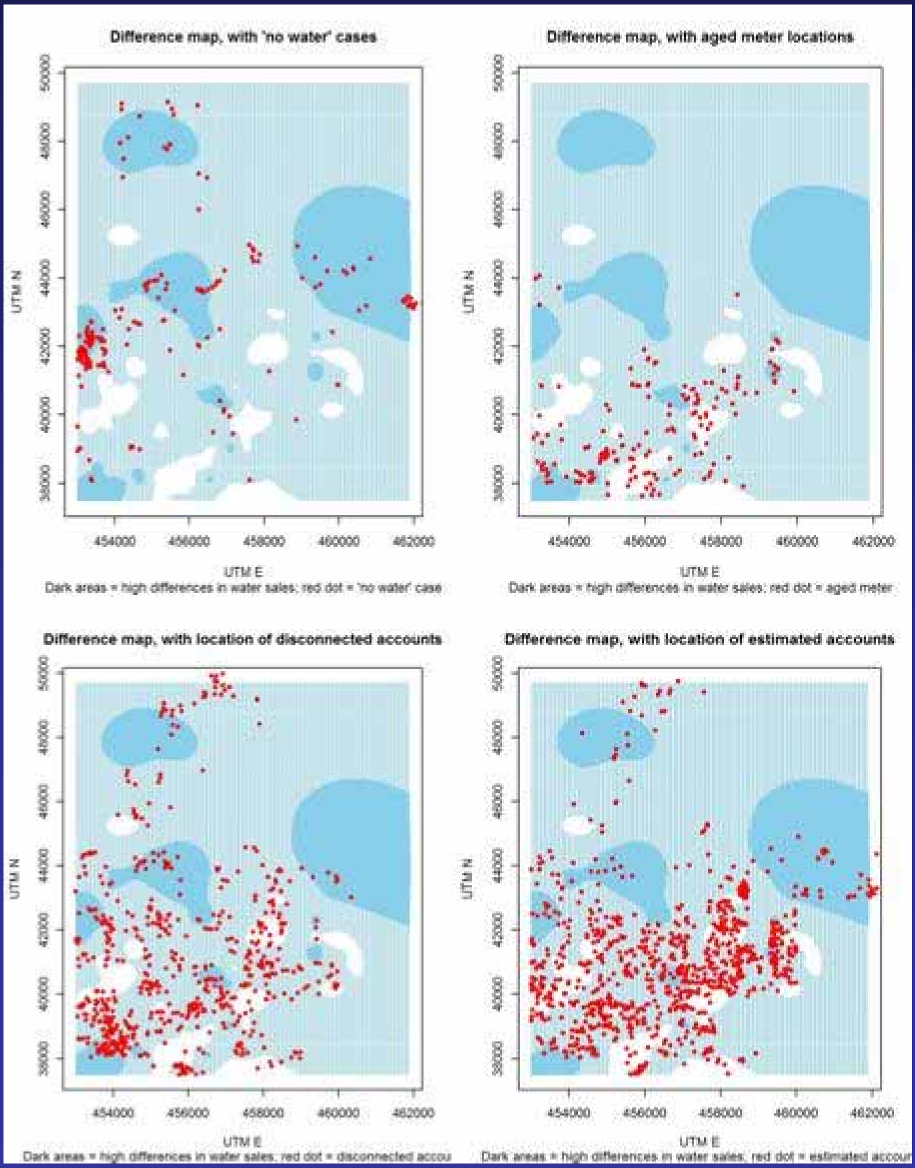
GEOSTATISTICAL ANALYSES:

(6) Difference map, highlighting areas with declining water sales




```
# Compare 2003 predicted values and 2007 predicted values and generate a difference map

str(k.grid)

str(kComp.grid)

diff <- (k.grid@data$var1.pred - kComp.grid@data$var1.pred)

str(diff)

diff <- as.data.frame(diff)

str(diff)

coordinates(diff) <- coordinates(d200312.grid)

str(diff)

print(spplot(diff, zcol = "diff", cuts = quantile(diff$diff, seq(0, 1, by = 0.1)), pretty = T, contour = T, col.regions = topo.colors(64), main = "Difference in log10 water sales volumes, 2003 and 2007", sub = log10Volume(2003) - log10 Volume(2007)", xlab = "UTM E", ylab = "UTM N", scales = list(draw = T), key.space = "right"))
```

GEOSTATISTICAL ANALYSES: (7) Overlays of reported field anomalies




```
# Load datasets of anomalies

NoWater <- read.table("NoWater.dat", header = T)

coordinates(NoWater) <- c("XCoord", "YCoord")

AgedMeter <- read.table("AgedMeter.dat", header = T)

coordinates(AgedMeter) <- c("XCoord", "YCoord")

Disconnected <- read.table("Disconnected.dat", header = T)

coordinates(Disconnected) <- c("XCoord", "YCoord")

Estimated <- read.table("Estimated.dat", header = T)

coordinates(Estimated) <- c("XCoord", "YCoord")

# Postplots of anomalies, superimposed on difference map

#(a) No water cases

plot(coordinates(diff), cex = 1.5*(diff$diff)/max(diff$diff), pch = 20, col =
"skyblue", main = "Difference map, with 'no water' cases", sub = high
differences in water sales; red dot = 'no water' case", xlab = "UTM E",
ylab = "UTM N")

points(coordinates(NoWater), col = "red", pch = 20, cex = 1.3)

#(b) Aged meters

plot(coordinates(diff), cex = 1.5*(diff$diff)/max(diff$diff), pch = 20, col =
"skyblue", main = "Difference map, with aged meter locations", sub =
areas = high differences in water sales; red dot = aged meter", xlab =
"UTM E", ylab = "UTM N")

points(coordinates(AgedMeter), col = "red", pch = 20, cex = 1.3)
```

RESULTS AND CONCLUSIONS:

- Out of eight factors investigated, “intermittent supply” cases correlated most with areas of declining water sales.
- Aged meters were also seen to have a contributing effect to low water sales.
- Disconnected and inactive accounts were numerous and widespread throughout the study area.
- Leaks and illegal cases seemed to have been too few to impact significantly on water sales.

ACKNOWLEDGMENTS

This study was carried out within the framework of the Sida/SAREC-funded project:

“Sustainable Technological Development in the Lake Victoria Basin”


Thank You!
Asanteni!
TACK!!