Lab03: Regression Kriging

# Part I: Trend Surface Model (8 points)

Task 01: You may use the script **SampleByClick.r** to pick manually in total 100 sample points. Alternatively, use your favorite GIS program. This task requires careful planning and perhaps a ***nested sampling strategy***. The predictive quality of your model very much depends on the selected sample points. (2 points)

Clearly justify your selection strategies of sample points based on the criteria listed below: (2 points)

(i) You want to avoid any bias in the predicted surface. Therefore, the average predicted elevations should match closely the average observed elevations in the study area. *How can you try to avoid this potential bias?*

Points were generally placed at an even density (though separated by varying distances) without concentrating too many toward any particular side of the study area. This should achieve a reasonably representative sampling of the actual elevations throughout the study area.

(ii) The extrapolation problem should be avoided and the prediction error, in particular at the edges of the study area, need to be minimized. *How many sample points should be assign to control for this problem?*

Ten points were assigned for placement around the study area edges. One point was placed in each corner and the others were spaced out around the edges.

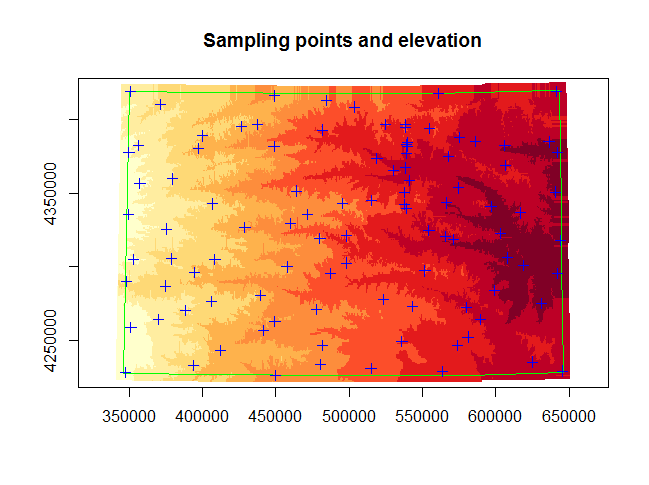
(iii) The rapid topographic variation along a transection of the river valleys and ridges needs to be captured properly. *How many sample points should be assigned to model this variability, where should they be placed and which variable in the data set measures it?*

A representative transection was selected and ten points were specifically placed at close intervals along the gradient between a ridge and a valley bottom.

(iv) In order to build a well-defined variogram all spatial scales of the inter-sample point distances need to be represented. *How many sample points should be assign to fill in missing distance ranges and where should these points be placed?*

After placing 70 points throughout the study area, 10 points around the edges, and 10 points across a transection, the remaining 10 points were placed at relatively close distances as needed since there were already many points representing the larger distances.

Finally, show the map of your sampling points and the given elevations similar to the map shown above.



Task 02: Estimate the 1st, 2nd and 3rd order trend-surface models. Include the distance to the nearest rivers as covariable. (1 point)

## First order

> polyForm1 <- makeTrendPolyForm(ELEVATION~RIVERDIST, ~X+Y,

shpDf=statPts, polyDeg=1)

> lmTrend1 <- lm(polyForm1, data=statPts)

> summary(lmTrend1)

Call:

lm(formula = polyForm1, data = statPts)

Residuals:

Min 1Q Median 3Q Max

-66.85 -21.45 -3.80 21.54 84.70

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 574.127 5.079 113.032 < 2e-16 \*\*\*

RIVERDIST 7.111 1.322 5.377 5.31e-07 \*\*\*

I(X^1 \* Y^0) -133.072 3.485 -38.188 < 2e-16 \*\*\*

I(X^0 \* Y^1) -7.616 3.281 -2.321 0.0224 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.72 on 96 degrees of freedom

Multiple R-squared: 0.9543, Adjusted R-squared: 0.9529

F-statistic: 668.3 on 3 and 96 DF, p-value: < 2.2e-16

## Second order

> polyForm2 <- makeTrendPolyForm(ELEVATION~RIVERDIST, ~X+Y,

shpDf=statPts, polyDeg=2)

> lmTrend2 <- lm(polyForm2, data=statPts)

> summary(lmTrend2)

Call:

lm(formula = polyForm2, data = statPts)

Residuals:

Min 1Q Median 3Q Max

-51.339 -16.962 -0.152 18.718 59.724

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 555.674 5.993 92.721 < 2e-16 \*\*\*

RIVERDIST 6.117 1.207 5.069 2.03e-06 \*\*\*

I(X^1 \* Y^0) -131.361 3.067 -42.824 < 2e-16 \*\*\*

I(X^0 \* Y^1) -6.289 2.933 -2.145 0.0346 \*

I(X^2 \* Y^0) 15.112 3.118 4.847 5.00e-06 \*\*\*

I(X^1 \* Y^1) -4.346 2.790 -1.558 0.1226

I(X^0 \* Y^2) 6.782 3.060 2.216 0.0291 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.7 on 93 degrees of freedom

Multiple R-squared: 0.9662, Adjusted R-squared: 0.9641

F-statistic: 443.7 on 6 and 93 DF, p-value: < 2.2e-16

Residual standard error: 27.69 on 93 degrees of freedom

Multiple R-squared: 0.9663, Adjusted R-squared: 0.9641

F-statistic: 444 on 6 and 93 DF, p-value: < 2.2e-16

## Third order

> polyForm3 <- makeTrendPolyForm(ELEVATION~RIVERDIST, ~X+Y,

shpDf=statPts, polyDeg=3)

> lmTrend3 <- lm(polyForm3, data=statPts)

> summary(lmTrend3)

Call:

lm(formula = polyForm3, data = statPts)

Residuals:

Min 1Q Median 3Q Max

-48.947 -15.907 -2.496 14.353 60.937

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 555.086 5.605 99.042 < 2e-16 \*\*\*

RIVERDIST 6.686 1.176 5.685 1.64e-07 \*\*\*

I(X^1 \* Y^0) -152.630 6.937 -22.001 < 2e-16 \*\*\*

I(X^0 \* Y^1) -11.726 6.627 -1.769 0.08024 .

I(X^2 \* Y^0) 15.438 2.911 5.303 8.24e-07 \*\*\*

I(X^1 \* Y^1) -2.343 2.685 -0.873 0.38523

I(X^0 \* Y^2) 7.530 2.990 2.518 0.01358 \*

I(X^3 \* Y^0) 7.249 3.373 2.149 0.03431 \*

I(X^2 \* Y^1) 3.339 2.712 1.231 0.22155

I(X^1 \* Y^2) 8.482 2.721 3.117 0.00246 \*\*

I(X^0 \* Y^3) 1.797 3.098 0.580 0.56343

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.74 on 89 degrees of freedom

Multiple R-squared: 0.9721, Adjusted R-squared: 0.969

F-statistic: 310 on 10 and 89 DF, p-value: < 2.2e-16

## Perform Predictions

> predTrend1 <- predict(lmTrend1, elevGrid, se.fit = TRUE)

> predTrend2 <- predict(lmTrend2, elevGrid, se.fit = TRUE)

> predTrend3 <- predict(lmTrend3, elevGrid, se.fit = TRUE)

Task 03: Map the three predicted trend-surfaces. Use a meaningful color ramp. (1 point)

> predFit1 <- predTrend1$fit

> predFit2 <- predTrend2$fit

> predFit3 <- predTrend3$fit

> n.col <- 9

> pal <- rev(brewer.pal(n.col,"YlOrRd"))

> predClass1 <- classInt::classIntervals(predFit1, n.col, style="equal")

> predClass2 <- classInt::classIntervals(predFit2, n.col, style="equal")

> predClass3 <- classInt::classIntervals(predFit3, n.col, style="equal")

> predCol1 <- classInt::findColours(predClass1,pal)

> predCol2 <- classInt::findColours(predClass2,pal)

> predCol3 <- classInt::findColours(predClass3,pal)

|  |  |
| --- | --- |
|  |  |
|  |  |

Task 04: Decide with the partial *F*-test, which of the three surface models is most appropriate for your given sample points. Interpreted the selected trend-surface regression model. (1 point)

> anova(lmTrend1, lmTrend2)

Analysis of Variance Table

Model 1: ELEVATION ~ RIVERDIST + I(X^1 \* Y^0) + I(X^0 \* Y^1)

Model 2: ELEVATION ~ RIVERDIST + I(X^1 \* Y^0) + I(X^0 \* Y^1) + I(X^2 \*

Y^0) + I(X^1 \* Y^1) + I(X^0 \* Y^2)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 96 96586

2 93 71337 3 25249 10.972 3.1e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The partial *F*-test with the first and second order trend surfaces is highly significant, indicating that the additional parameters in the second order model provide a significant improvement over the first order.

> anova(lmTrend2, lmTrend3)

Analysis of Variance Table

Model 1: ELEVATION ~ RIVERDIST + I(X^1 \* Y^0) + I(X^0 \* Y^1) + I(X^2 \*

Y^0) + I(X^1 \* Y^1) + I(X^0 \* Y^2)

Model 2: ELEVATION ~ RIVERDIST + I(X^1 \* Y^0) + I(X^0 \* Y^1) + I(X^2 \*

Y^0) + I(X^1 \* Y^1) + I(X^0 \* Y^2) + I(X^3 \* Y^0) + I(X^2 \*

Y^1) + I(X^1 \* Y^2) + I(X^0 \* Y^3)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 93 71337

2 89 58988 4 12350 4.6583 0.001845 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The *F*-test for the second and third order models is also significant, which suggests that the third order trend surface is the best of the three.

For the third order model, the distance to nearest river is highly significant and has a positive value, meaning that greater distances from rivers are associated with higher elevations, which is as expected. The first order longitude term is very significant with a large negative value, indicating that elevation generally is lower moving toward the right side of the study area. The first order latitude term is not significant, which is sensible because the elevation visually doesn’t have much average difference between the top and bottom portions of the map.

Task 05: Evaluate the prediction quality of your most appropriate trend surface model. Does the histogram of observed elevations match that based on the predicted values? Does your prediction model lead to biased overall elevation estimates? If yes, what may be the cause? (1 point)

> hist(elevGrid$ELEVATION)

> hist(predTrend3$fit)

|  |  |
| --- | --- |
|  |  |

The overall shape of the histograms are similar, but the trend model somewhat overpredicts at the lower elevations, such as in the 400–450 category. It is possible that the points assigned to measure the transect were in a portion of the study area that has overall lower than average elevations, so having that cluster of points in that area could have biased the prediction even though the points within the transect capture a wide range of elevations. However, the mean elevations of the elevation grid and sample points are similar (603.3 for the grid and 595.5 for the sample points).

Task 06: For your most appropriate model, map the ***standard errors*** of the prediction surfaces. Use a meaningful color ramp. Interpret the general pattern in the standard errors. In particular evaluate the standard errors at the edges of the study area relative to those in the center? (1 point)

> predSe <- predTrend3$se

> n.col <- 9

> pal <- brewer.pal(n.col,"Reds")

> seClass <- classInt::classIntervals(predSe, n.col, style="equal")

> seCol <- classInt::findColours(seClass,pal)

> plot(elevGrid,axes=T,col=seCol,pch=15,cex=2,

xlim=boxPred[1,],ylim=boxPred[2,],

main="Standard Error: 3rd Order Trendsurface")

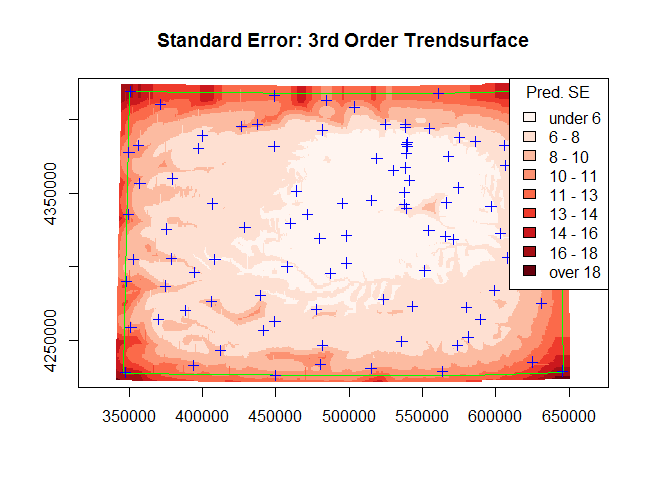
> plot(landBnd,axes=T,border="green",add=T)

> plot(statPts,col="blue",add=T)

> legend("topright", title = "Pred. SE",

legend = leglabs(round(seClass$brks, digits = 0)), fill = pal,

bty = "o", ncol = 1)



The standard errors are lower toward the middle of the study area and higher toward the edges. Predictions at the edges are less certain because more extrapolation is required. The standard errors are also lower near large groups of sample points.

Task 07: For your most appropriate model, calculate the error component (residual surface: observed DEM minus predicted trend DEM). Map this pattern with a bipolar map theme (zero is the neutral value) and overlay the river network onto your residual map. Interpret this residual pattern. (1 point)

> elevGrid$res <- elevGrid$ELEVATION - predTrend3$fit

> n.col <- 9

> pal <- rev(brewer.pal(n.col,"RdBu"))

> resClass <- classInt::classIntervals(elevGrid$res, n.col, style="equal")

> resCol <- classInt::findColours(resClass,pal)

> plot(elevGrid,axes=T,col=resCol,pch=15,cex=2,

xlim=boxPred[1,],ylim=boxPred[2,],

main="Residuals: 3rd Order Trendsurface")

> plot(landBnd,axes=T,border="green",add=T)

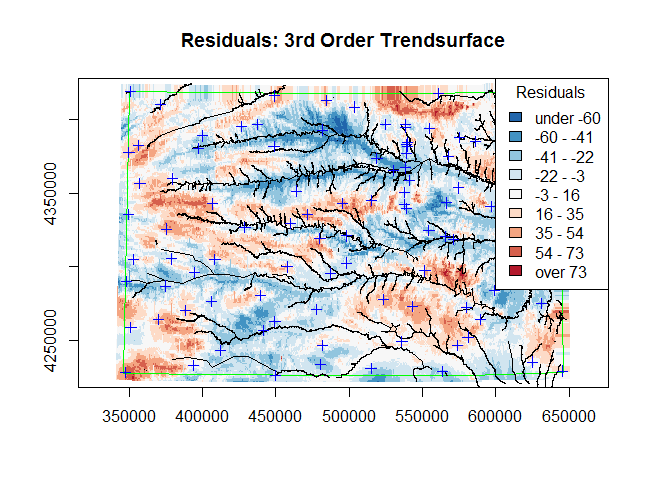
> plot(riverShp,axes=T,col="black",add=T)

> plot(statPts,col="blue",add=T)

> legend("topright", title = "Residuals",

legend = leglabs(round(resClass$brks, digits = 0)), fill = pal,

bty = "o", ncol = 1)



In general, there appears to be little to no east/west trend in the residuals. However, the trend model was much less effective at capturing the finer scale undulation associated with the rivers—elevations near rivers were consistently overpredicted and areas farther away were often underpredicted.

# Part II: Variogram Estimation (2 points)

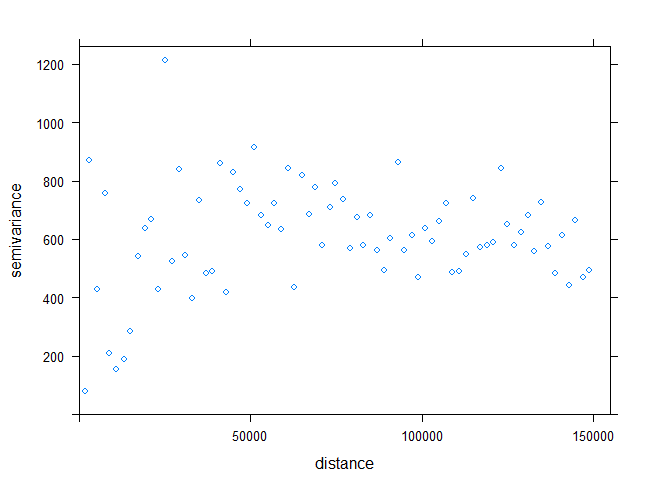
Task 08: Estimate the variogram function based on the error component at the sampling locations from Part I. Show the necessary plots and interpret them by exploring possible anisotropy, range, sill and nugget effects.

> statPts$res <- residuals(lmTrend3)

> gstat1 <- gstat(id="elevTrend", formula=res~X+Y, data=statPts)

> vgm1 <- variogram(gstat1, cutoff=150000, width=2000)

> plot(vgm1)



The curve envisioned through the middle of the points appears to level off at a distance of approximately 40,000 with a semivariance of about 650, so these will be used as the estimated range and sill values. The nugget will be estimated at 100 because there are no distances with a semivariance lower than about 100.

# Part III: Kriging Interpolation of the error component (3 points)

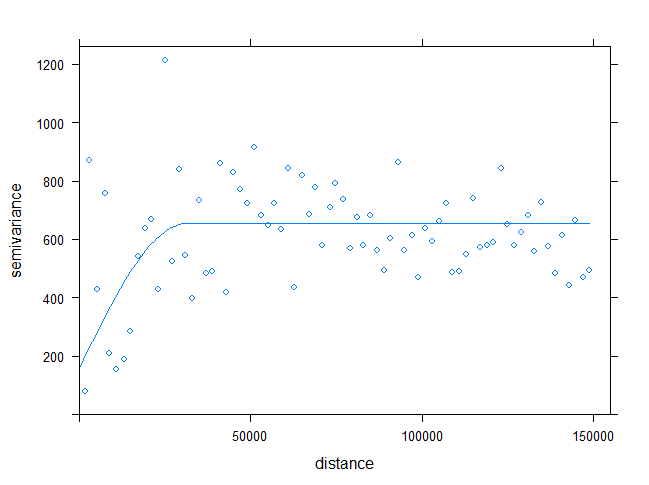
Task 09: Predict the error component by Kriging for all locations. Justify your choice of the Kriging model. Map the surface of the predicted error component with an appropriate color ramp. (1.5 point)

The spherical model was chosen based on the overall shape of the semivariance point distribution, where the first portion below the range appears sharply distinct from the portion beyond the range – there does not appear to be the smooth curve that the exponential model would be a good fit for.

> fit1 <- fit.variogram(vgm1, model=vgm(psill=650, model="Sph",

range=40000, nugget=100))

> plot(vgm1, model=fit1)



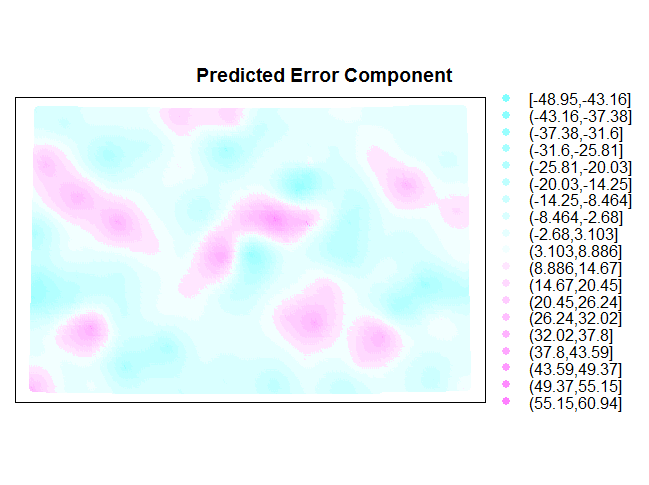
> gstat1 <- gstat(gstat1, id="elevTrend", beta=0, model=fit1)

> pred1 <- predict(gstat1, newdata=elevGrid)

> spplot(pred1,zcol="elevTrend.pred",col.regions=cm.colors(20),

cuts=19,main="Predicted error component",

key.space="right")



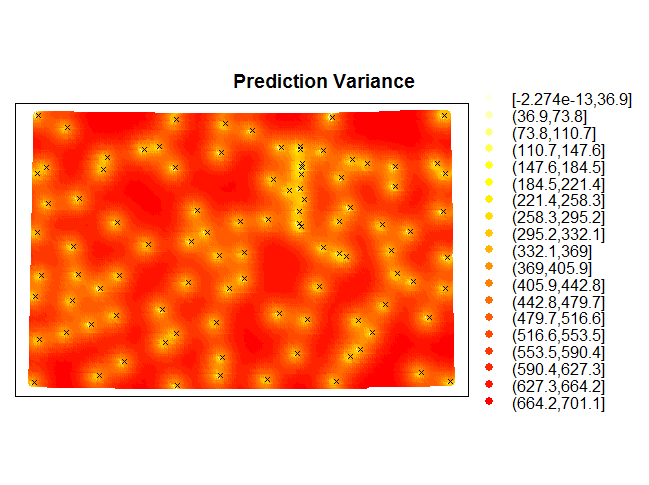
Task 10: Estimate the uncertainty of the error component for all locations. Map the uncertainty surface with an appropriate color ramp. (1.5 point)

> pts <- list("sp.points", statPts, pch=4, col="black", cex=0.5)

> spplot(pred1,zcol="elevTrend.var",col.regions=rev(heat.colors(20)),

cuts=19,sp.layout=list(pts),main="Prediction Variance",

key.space="right")



As expected, the error component prediction uncertainty very consistently increases as the distance from sample points increases.

# Part IV: Combining Frist and Second Order Components (3 points)

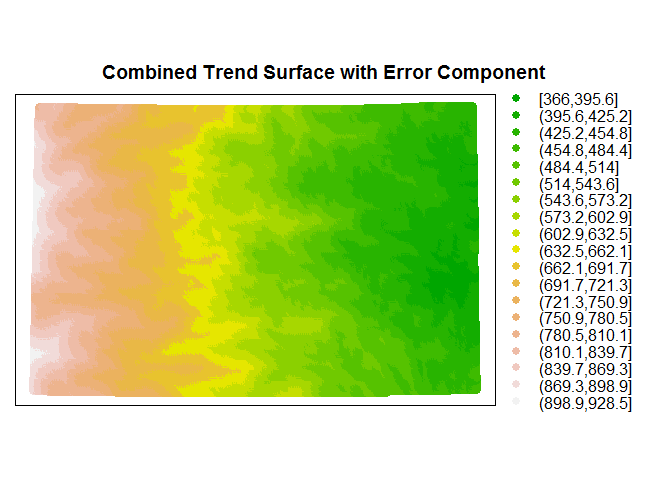
Task 11: Combine the predicted trend-surface with the predicted error component to obtain the overall predicted DEM surface. Map this predicted surface with a proper color ramp. (1 point)

> elevGrid$dem <- pred1$elevTrend.pred + predTrend3$fit

> spplot(elevGrid, zcol="dem",col.regions=terrain.colors(20),cuts=19,

pretty=TRUE,main="Combined Trend Surface with Error Component",

key.space="right")



Task 12: Combined the trend-surface prediction uncertainty with the kriging uncertainty in the standard deviation scale. Map the uncertainty surface with a proper color ramp. (1 point)

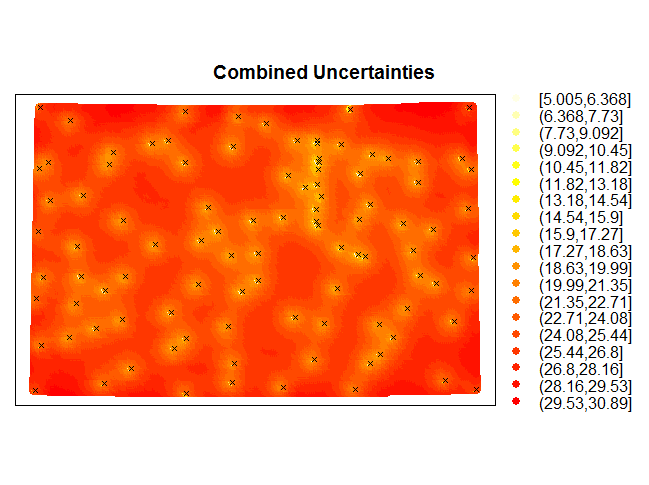
> elevGrid$uncertainty <- sqrt(pred1$elevTrend.var + predTrend3$se.fit^2)

> spplot(elevGrid,zcol="uncertainty",col.regions=rev(heat.colors(20)),

cuts=19,sp.layout=list(pts),contour=TRUE,labels=FALSE,

pretty=TRUE,col="brown",main="Combined Uncertainties",

key.space="right")



Task 13: Calculate the root mean squared error of your overall predicted DEM values by comparing it against the observed DEM value of the Kansas topographic surface. (1 point)

> (RMSE <- sqrt(mean((elevGrid$dem - elevGrid$ELEVATION)^2)))

[1] 21.50603