Applied geostatistics
Exercise 5a: Predicting from point samples (Part 3)
Using secondary information

D G Rossiter
University of Twente, Faculty of Geo-Information Science & Earth Observation (ITC)

June 26, 2010

Contents

1 Introduction **1**

2 Setup **1**

3 Feature-space modelling **2**
3.1 Answers for feature-space modelling 6

4 Computing and modelling the residual variogram **7**
4.1 Answers for Computing and modelling the residual variogram 11

5 Creating the interpolation grid **11**
5.1 Answers for Creating the interpolation grid 13

6 Kriging with external drift **14**
6.1 Answers for Kriging with external drift 20

7 Validation **21**

8 Cleaning up **21**

9 Self-test **21**

References **23**

Version 1.51. Copyright © 2007–2008 ITC All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (http://www.itc.nl/personal/rossiter).

Index of R concepts 24

1 Introduction

In Exercise 4 §3 we saw how a feature-space attribute could be used for prediction. This had the disadvantage that the prediction was the same everywhere in a stratum, and did not take account of local spatial dependence. In Exercise 4 §4 we saw how to predict using Ordinary Kriging; this used a model of local spatial dependence but no information on feature-space attributes other than the target variable. These two approaches can be combined into a **mixed predictor** that uses both sources of information. There are three main approaches for this:

1. Kriging with External Drift (KED): model residual spatial variability, use classes and this model with all samples to predict;
2. Regression Kriging (RK): compute feature-space model, compute residuals from this, model residual spatial variability, use Simple Kriging (with expected value 0) to predict residuals, add in the class means;
3. Stratified Kriging (StK): stratify the area by class, model the variogram in each class separately, predict by OK in each class separately, merge the classes for the final map.

These each have advantages and disadvantages:

1. KED models both local and feature-space in one kriging system, so it is the easiest extension to OK. The mathematics are exactly the same as for UK;
2. RK is quite flexible but requires that we model the feature space and residuals separately;
3. StK allows the possibility of different spatial structure (e.g. variograms) in the different strata; however the sample size must be large enough to estimate this structure in each stratum separately.

In this exercise we will examine KED.

After completing this exercise you should be able to:

1. Model feature-space relations between a continuous target variable and a categorical (classified) predictor variable;
2. Model local spatial dependence after accounting for these feature-space relations;
3. Predict attributes over a region using Kriging with External Drift.

2 Setup

We continue with the Jura dataset from Exercise 4.

Task 1 : Set up this exercise as explained in the following list. •

1. If R is not already running, start it. If you haven't already done so, load the `gstat` and `sp` libraries.
2. If the calibration dataset `jura.cal` is not loaded as a spatial object, do so.
3. If the fitted variogram model `vmf` from Exercise 4 §4.1 is no longer in the workspace, re-create it.

If you followed the instructions in Exercise 4, these should all have been saved in file `JuraEx4.RData`, so they can be restored with the `load` method:

```
> load("JuraEx4.RData")
```

> **Note:** If this file is not in the current directory, you either have to change the directory with the `setwd` method, or add the full path to the argument of the `load` method.

3 Feature-space modelling

The first step in a mixed predictor is to see whether the feature space provides any useful information with respect to the target variable.

Q1 : *What are the categorical attributes in the Jura dataset? Are these expected to be predictors of heavy metals in soils? Why or why not?* *Jump to A1* •

```
> str(jura.cal@data)
```

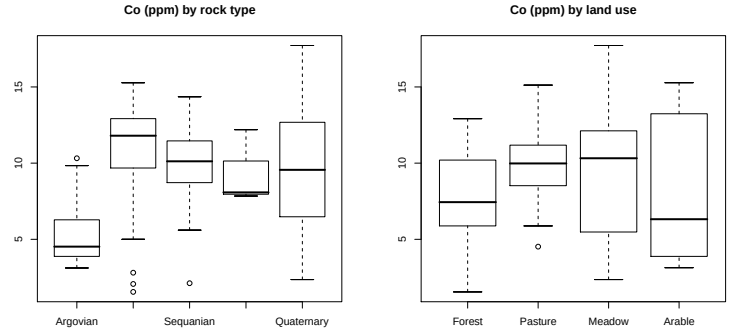
```
'data.frame':        259 obs. of  9 variables:
 $ Rock: Factor w/ 5 levels "Argovian","Kimmeridgian",..: 3 2 3 2 5 5 5 1 1 3 ...
 $ Land: Factor w/ 4 levels "Forest","Pasture",..: 3 2 2 3 3 3 3 3 3 3 ...
 $ Cd  : num  1.74 1.33 1.61 2.15 1.56 ...
 $ Cu  : num  25.72 24.76  8.88 22.70 34.32 ...
 $ Pb  : num  77.4 77.9 30.8 56.4 66.4 ...
 $ Co  : num   9.32 10.00 10.60 11.92 16.32 ...
 $ Cr  : num  38.3 40.2 47.0 43.5 38.5 ...
 $ Ni  : num  21.3 29.7 21.4 29.7 26.2 ...
 $ Zn  : num  92.6 73.6 64.8 90.0 88.4 ...
```

We first **visualize** whether there is any difference in the target variable Co (cobalt) due to these factors.

Task 2 : Display **classified boxplots** of Co by rock type and by land use. •

We use the `boxplot` method with a formula:

```
> par(mfrow = c(1, 2))
> boxplot(Co ~ Rock, data = jura.cal@data, main = "Co (ppm) by rock type")
> boxplot(Co ~ Land, data = jura.cal@data, main = "Co (ppm) by land use")
> par(mfrow = c(1, 1))
```



Q2 : *Do the areas with different **rock types** have similar distributions of Co? If not, what are the major differences?* *Jump to A2* •

Q3 : *Do the areas with different **land uses** have similar distributions of Co? If not, what are the major differences?* *Jump to A3* •

Q4 : *Which of the two classified variables is expected to provide more differentiation for this target variable?* *Jump to A4* •

After visualizing, we can verify our impressions by a **linear model** of the target variable as modelled by the categorical variable; this one-way Analysis of Variance (ANOVA) is computed with the `lm` method and summarized with `summary.lm` (which is automatically called by the generic `summary` method for objects of class `lm`).

Task 3 : Compute and summarize linear models of Co predicted by rock type and land use, separately. •

Note that we apply the `summary` method to a linear model object, created by the `lm` method. The linear model object is saved in the workspace with the `<-` assignment operator.

```
> summary(mr <- lm(Co ~ Rock, data = jura.cal@data))

Call:
lm(formula = Co ~ Rock, data = jura.cal@data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4946 -1.5642  0.0868  1.8800  8.1245

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.393      0.404   13.34  < 2e-16 ***
RockKimmeridgian   5.651      0.515   10.97  < 2e-16 ***
RockSequanian      4.582      0.549    8.35  4.4e-15 ***
RockPortlandian    3.980      1.747    2.28    0.024 *
```

```
RockQuaternary      4.202      0.567    7.42  1.8e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.94 on 254 degrees of freedom
Multiple R-Squared: 0.334,        Adjusted R-squared: 0.323
F-statistic: 31.8 on 4 and 254 DF,  p-value: <2e-16
```

```
> summary(ml <- lm(Co ~ Land, data = jura.cal@data))

Call:
lm(formula = Co ~ Land, data = jura.cal@data)

Residuals:
   Min     1Q Median     3Q    Max
-7.034 -2.698  0.286  2.571  8.326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.683      0.614   12.51   <2e-16 ***
LandPasture    2.381      0.775    3.07   0.0023 **
LandMeadow     1.710      0.673    2.54   0.0116 *
LandArable     0.689      1.694    0.41   0.6847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.53 on 255 degrees of freedom
Multiple R-Squared: 0.0378,        Adjusted R-squared: 0.0265
F-statistic: 3.34 on 3 and 255 DF,  p-value: 0.02
```

Q5 : *How much of the variability is explained by each model? (Look at the adjusted R^2) Which is thus the better predictor?* *Jump to A5* •

Perhaps there is an interaction between the two predictors.

Task 4 : Compute and summarize a linear model of Co predicted by rock type and land use, together, both as **additive** factors (no interaction) and as **interacting** factors. •

The formula for additive models names the two factors, joined by `+`; the interaction model joins the names by `*`.

First, the additive model:

```
> summary(mrl <- lm(Co ~ Rock + Land, data = jura.cal@data))

Call:
lm(formula = Co ~ Rock + Land, data = jura.cal@data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5911 -1.8489 -0.0316  1.7507  8.0947

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.158      0.624    5.06  8.1e-07 ***
```

```
RockKimmeridgian    5.983      0.506   11.83  < 2e-16 ***
RockSequanian       4.400      0.545    8.07  2.9e-14 ***
RockPortlandian     4.503      1.693    2.66  0.00833 **
RockQuaternary      3.856      0.552    6.99  2.5e-11 ***
LandPasture         2.525      0.650    3.88  0.00013 ***
LandMeadow          2.611      0.572    4.56  8.0e-06 ***
LandArable          1.486      1.379    1.08  0.28241
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.84 on 251 degrees of freedom
Multiple R-Squared: 0.387,        Adjusted R-squared: 0.37
F-statistic: 22.6 on 7 and 251 DF,  p-value: <2e-16
```

Then, the model with interactions:

```
> summary(mrli <- lm(Co ~ Rock * Land, data = jura.cal@data))

Call:
lm(formula = Co ~ Rock * Land, data = jura.cal@data)

Residuals:
      Min        1Q    Median        3Q       Max
-7.64e+00 -1.24e+00 -9.90e-16  1.48e+00  7.72e+00

Coefficients: (2 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.806      1.014    5.72  3.1e-08 ***
RockKimmeridgian               2.161      1.165    1.86  0.06472 .
RockSequanian                  2.674      1.852    1.44  0.15002
RockPortlandian                6.394      2.869    2.23  0.02675 *
RockQuaternary                -0.740      3.795   -0.19  0.84558
LandPasture                    2.701      1.493    1.81  0.07170 .
LandMeadow                    -0.927      1.102   -0.84  0.40104
LandArable                    -1.926      2.869   -0.67  0.50273
RockKimmeridgian:LandPasture   1.337      1.720    0.78  0.43772
RockSequanian:LandPasture     -1.416      2.218   -0.64  0.52371
RockPortlandian:LandPasture   -7.061      4.078   -1.73  0.08468 .
RockQuaternary:LandPasture    -0.347      4.099   -0.08  0.93268
RockKimmeridgian:LandMeadow    5.054      1.306    3.87  0.00014 ***
RockSequanian:LandMeadow       2.729      1.958    1.39  0.16458
RockPortlandian:LandMeadow    -3.193      3.952   -0.81  0.41988
RockQuaternary:LandMeadow      5.863      3.839    1.53  0.12804
RockKimmeridgian:LandArable    9.239      3.970    2.33  0.02078 *
RockSequanian:LandArable       3.226      3.773    0.86  0.39339
RockPortlandian:LandArable        NA         NA      NA       NA
RockQuaternary:LandArable         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.68 on 241 degrees of freedom
Multiple R-Squared: 0.474,        Adjusted R-squared: 0.437
F-statistic: 12.8 on 17 and 241 DF,  p-value: <2e-16
```

Q6 : *In the second model there is one highly-significant* ***interaction*** *between the two factors. Which levels of these factors? What is the interpretation of this interaction?* *Jump to A6*

•

Another way to compare **hierarchical** feature-space models (i.e. where they are increasingly-complex) is ANOVA. This shows how much more variance is explaained by the more complex models, and whether this is a significant difference or could have occurred by chance.

Task 5 : Compare the interactive, additive, and single-factor (rock type) models with an ANOVA. •

We use the `anova` method:

```
> anova(mrli, mrl, mr)

Analysis of Variance Table

Model 1: Co ~ Rock * Land
Model 2: Co ~ Rock + Land
Model 3: Co ~ Rock
  Res.Df  RSS  Df Sum of Sq    F  Pr(>F)
1    241 1736
2    251 2023 -10      -288 3.99 4.5e-05 ***
3    254 2200  -3      -177 8.18 3.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q7 : *Are the models significantly different? Which increase in complexity gives the most improvement?* *Jump to A7* •

Q8 : *In words, what does it mean that the additive and interactive models are significantly better than the single-factor model?* *Jump to A8* •

We have seen that the interactive model of both rock type and land use gives the best results; however, for the rest of this exercise we will use only the single-factor model of rock type; this will allow us to visualize the differences in predictions against a single map (i.e. that of rock type) and should help in understanding. In practice you would probably use the best model, since it is quite a bit better than either single-factor model.

3.1 Answers for feature-space modelling

A1 : *Rock type (field `Rock`) and land use (field `Land`). Both might be associated with soil metals: different rocks have different minerals, which have different proportions of various metals; different land uses may be treated differently (e.g. fertilization, removal of plant material) which may cause differential concentrations of metals in the soil.* *Return to Q1* •

A2 : *The Argovian rocks have a distribution of Co values that is well below the others (although there is some overlap). The Quaternary sediments have the*

complete range (probably because they are derived from all the others). *Return to Q2* •

A3 : *There is not much difference between land uses, except that pasture has no Co values less than 5 mg kg^{-1}and meadow has all the values greater than 15 mg kg^{-1}.* *Return to Q3* •

A4 : *The rock types appear more discriminating.* *Return to Q4* •

A5 : *Adjusted R^2 are 32.3% (rock type) and 2.7% (land use); clearly rock type is the better predictor.* *Return to Q5* •

A6 : *The interaction between Kimmeridgian rocks and Meadow land use is very highly significant. The estimate is positive (5.054); this means that Co concentrations in meadows with soils developed on these rocks is very significantly higher (by 5 mg kg^{-1}) than would be expected from the two factors taken independently. There is some positive **synergy** between this rock type and this land use; the physical reason for this is not clear.* *Return to Q6* •

A7 : *Both improvements are highly significant (probability $< 3.310^{-5}$ and $< 4.510^{-5}$, respectively, that the improvement is due to chance. The improvement from the additive to the interactive model is larger (see the differences in mean sum of squares).* *Return to Q7* •

A8 : *Additive: both factors are significant predictors of Co; Interactive: some combinations of rock type and land use have significantly different Co concentrations than would be expected from the addition of the effects of the two factors separately.* *Return to Q8* •

4 Computing and modelling the residual variogram

We have seen that the feature space explains a fair amount of the variation. But we also saw that the maximum explained was under half of the total. Where is the rest?

One possibility is the local spatial structure, as we saw for OK (Exercise 4 §4). Now the question is, **after** accounting for the feature-space predictor, is there still local structure?

This is shown by the **residual variogram**, i.e. a plot of the semi-variances of **residuals** (not original values) against separation of point pairs. Fortunately, `gstat` computes these directly, if you provide an appropriate model formula; you do not have to compute the residuals manually.

Task 6 : Compute and display the residual variogram for Co, accounting

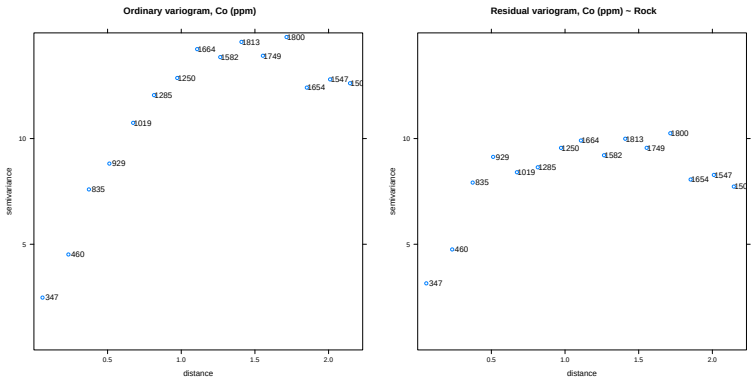
for the feature-space predictor of rock type. Compare it with the ordinary variogram. •

We use the `variogram` method and specify the spatial dependence with the formula `Co ~ Rock` (as opposed to `Co ~ 1` in the ordinary variogram). This has the same meaning as in the `lm` model specification: the cobalt concentration is to be predicted from the rock type; then the residuals are to be modelled spatially.

```
> vr <- variogram(Co ~ Rock, loc = jura.cal)
> v <- variogram(Co ~ 1, loc = jura.cal)
```

We plot them on the same scale, using the optional `ylim` argument; the maximum is computed by taking the next highest integer (using the `ceiling` method) than the maximum of all the semivariances:

```
> gamma.max <- ceiling(max(v$gamma, vr$gamma))
> plot.1 <- plot(vr, ylim=c(0,gamma.max), pl=T,
+    main="Residual variogram, Co (ppm) ~ Rock")
> plot.2 <- plot(v, ylim=c(0,gamma.max), pl=T,
+    main="Ordinary variogram, Co (ppm)")
> print(plot.2, split=c(1,1,2,1), more=T)
> print(plot.1, split=c(2,1,2,1), more=F)
> rm(gamma.max, plot.1, plot.2)
```

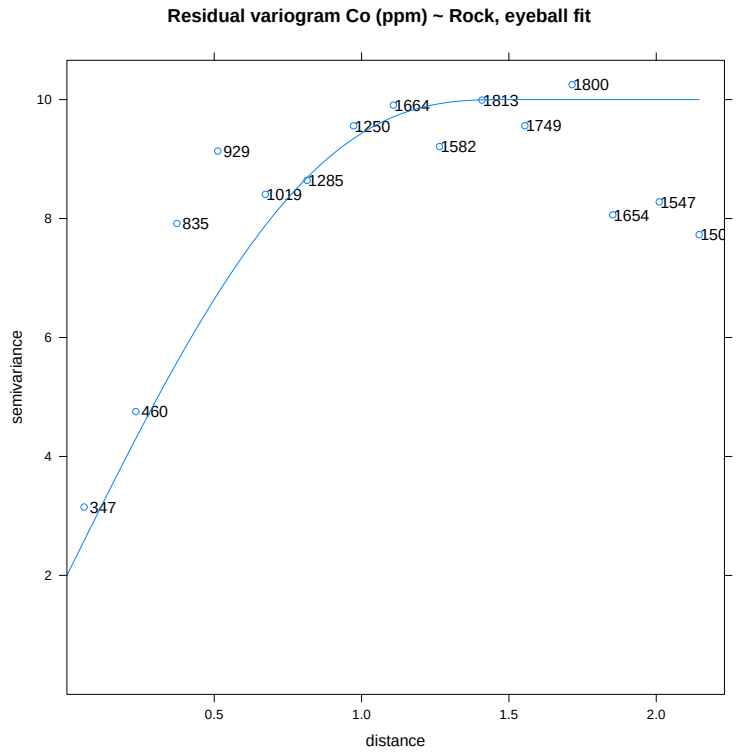


Q9 : *Describe the differences between the ordinary and residual variograms (model form, total sill, nugget, range).* *Jump to A9* •

Task 7 : Model the residual variogram; compare this to the model for the original variogram. •

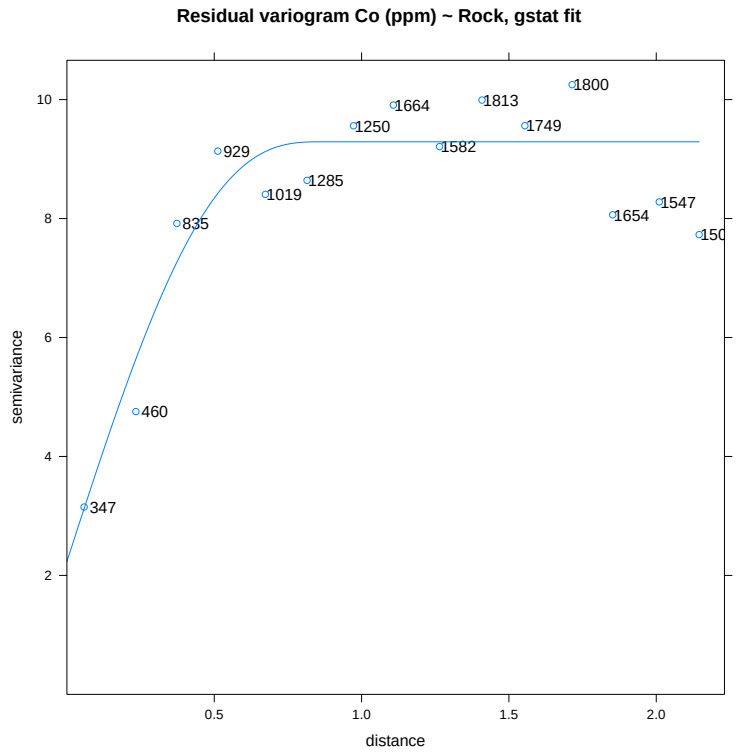
A pentaspherical model is indicated, since that model was with the original variable. We first make a guess with the `vgm` method:

```
> vrm <- vgm(8, "Pen", 1.5, 2)
> print(plot(vr, pl=T, model=vrm,
+       main="Residual variogram Co (ppm) ~ Rock, eyeball fit"))
```



We then ask `gstat` to fit it with the `fit.variogram` method:

```
> (vrmf <- fit.variogram(vr, model=vrm))
> print(plot(vr, pl=T, model=vrmf,
+       main="Residual variogram Co (ppm) ~ Rock, gstat fit"))

  model  psill   range
1   Nug 2.2343 0.00000
2   Pen 7.0559 0.86662
```



Now we can compare the residual model to the original:

```
> vmf

  model   psill  range
1   Nug  1.3712 0.0000
2   Pen 12.9322 1.5239

> vrmf

  model  psill   range
1   Nug 2.2343 0.00000
2   Pen 7.0559 0.86662

> vmf$range[2] - vrmf$range[2]

[1] 0.65728

> 1 - (vmf$range[2] - vrmf$range[2])/vmf$range[2]

[1] 0.56869

> sum(vmf$psill)

[1] 14.303

> sum(vrmf$psill)

[1] 9.2903

> 1 - sum(vrmf$psill)/sum(vmf$psill)

[1] 0.35048
```

Q10 : *How much has the range been reduced? How much has the total sill been reduced?* *Jump to A10* •

Q11 : *In words, what is the difference between the models? How can you explain this?* *Jump to A11* •

4.1 Answers for Computing and modelling the residual variogram

A9 : *The model form looks more-or-less the same (so the pentaspherical model we used for OK can also be used for KED); the total sill is considerably lower (10 vs. 16 or so) for the residual variogram; the nugget is similar; the range is shorter (0.5 vs. 1.2 or so).* *Return to Q9* •

A10 : *The range has been reduced from 1.52 km to 0.87 km; that is 56% of the original. The total sill has been reduced from 14.3 mg kg^{-12}to 9.3 mg kg^{-12}; that is a 35% reduction in variability.* *Return to Q10* •

A11 : *The spatial dependence is more local (about half the range) and lower overall (about 1/3 lower). The rest of the apparent spatial structure has been accounted for in feature space.* *Return to Q11* •

5 Creating the interpolation grid

Now we want to predict by KED; but first we need an interpolation grid. We already have one which covers a rectangle enclosing the sample points, created in Exercise 4 §4.4. But, this only has point locations where we want to predict. For KED we also need to know the **feature-space** values of the categorical predictor (e.g. rock type and land use) at **every point** to be predicted.

Fortunately, such a grid has been prepared by Pierre Goovaerts, author of a well-known geostatistics text [1] which uses the dataset as its running example. He has kindly provided us this grid, which he created by manually overlaying a 50 m grid (at each intersection of a grid centred every 0.05 km in both directions) of prediction points over printed rock type and land use maps on a light table, and manually recording the category. To save you work, I have prepared this as an R spatial object, as file `Jura50.RData` which can be loaded with the `load` method as workspace object `jura50`.

> **Note:** Since these notes were first written, the same grid has been added to the sample `jura` dataset of the `gstat` package, as object `jura.grid`.

Task 8 : Load the file `Jura50.RData` and summarize the newly-created workspace object `jura50`. •

```
> load("Jura50.RData")
> summary(jura50)
```

Q12 : *Do the proportions of rock types in the grid match those in the calibration data set?* *Jump to A12* •

We summarize the counts of each class with the `table` method:

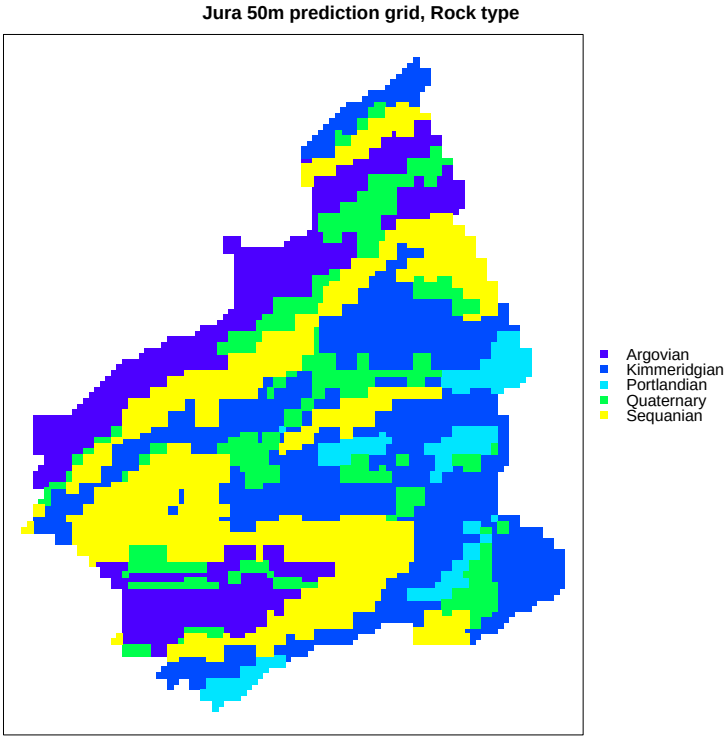
```
> table(jura50$Rock)/length(jura50$Rock)

    Argovian Kimmeridgian    Sequanian  Portlandian   Quaternary
    0.198926     0.341783     0.273292     0.053047     0.132953

> table(jura.cal$Rock)/length(jura.cal$Rock)

    Argovian Kimmeridgian    Sequanian  Portlandian   Quaternary
    0.204633     0.328185     0.243243     0.011583     0.212355
```

Task 9 : Display the distribution of rock types and land uses as maps. •

We use the `spplot` method, saving the two maps as lattice graphics objects for later use.

First the rock type:

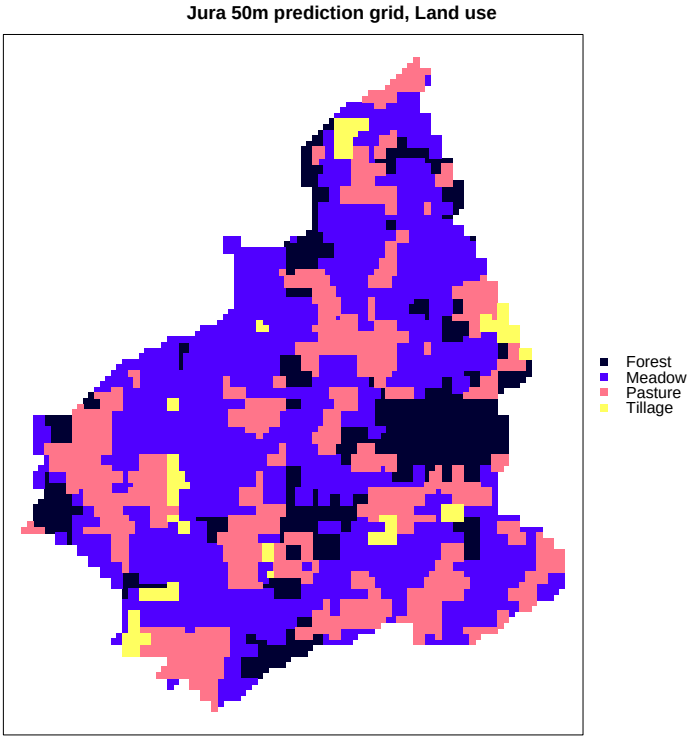
```
> map.rock <- spplot(jura50, zcol = "Rock", pch = 15, col.regions = topo.colors(5),
+     key.space = "right", main = "Jura 50m prediction grid, Rock type")
> print(map.rock)
```



Q13 : *What is the spatial pattern of the rock types?* *Jump to A13* •

Now the land use:

```
> map.land <- spplot(jura50, zcol = "Land", pch = 15, col.regions = bpy.colors(4),
+     key.space = "right", main = "Jura 50m prediction grid, Land use")
> print(map.land)
```



Q14 : *What is the spatial pattern of the land uses?* *Jump to A14* •

5.1 Answers for Creating the interpolation grid

A12 : *Quaternary rocks are substantially over-represented in the calibration set (21% vs. 14%); Portlandian rocks are substantially under-represented (1% vs. 5%); proportions of the other classes are about equal between the grid and calibration set.* *Return to Q12* •

A13 : *Rock types are generally outcropping following a SW-NE strike.* *Return to Q13* •

A14 : *Land uses form local patches of varying sizes. Most of the area is meadow; there are larger patches of pasture and smaller patches of forest and especially tillage. So most of the area is used for animal production (either natural meadows or managed pasture).* *Return to Q14* •

6 Kriging with external drift

We now have a **model of spatial dependence** of the residuals from rock type (model `vrmf`), and a prediction grid with the rock type known at each location (grid `jura50`). This is what is necessary for KED.

Task 10 : Predict the Co concentration over the 50 m grid, using kriging with external drift (KED), with rock type as the feature-space predictor. •

The `krige` method will do this, again with a formula `Co ~ Rock` that shows the feature-space dependence:

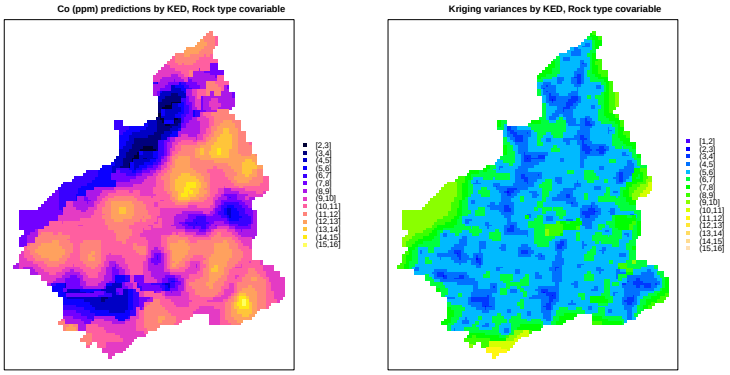
```
> kr.50 <- krige(Co ~ Rock, loc = jura.cal, newdata = jura50,
+     model = vrmf)

[using universal kriging]
```

! → **Note carefully**the use of the **residual** variogram model: `model=vrmf`.

> **Note:** You may have noticed that the `krige` method gives the message `[using universal kriging]` as it works; this is because the form of the kriging equations is exactly the same for KED as for what we have called UK in Exercise 5 §4. We have reserved the term "universal kriging" for the case where the co-predictors are the coördinates; other authors also include feature-space predictors in this term, so they do not use the term "kriging with external drift" at all.

Task 11 : View the predictions and the kriging variances. •

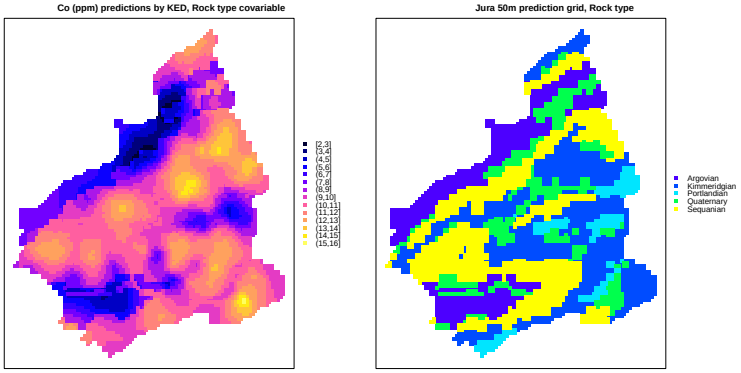
```
> plot.ked <- spplot(kr.50, zcol = "var1.pred", contour = T,
+     pch = 15, by = seq(2, 16, by = 1), cuts = seq(2,
+         16, by = 1), col.regions = bpy.colors(64), key.space = "right",
+     main = "Co (ppm) predictions by KED, Rock type covariable")
> plot.ked.v <- spplot(kr.50, zcol = "var1.var", contour = T,
+     pch = 15, by = seq(1, 16, by = 1), cuts = seq(1,
+         16, by = 1), col.regions = topo.colors(64), key.space = "right",
+     main = "Kriging variances by KED, Rock type covariable")
> print(plot.ked, split = c(1, 1, 2, 1), more = T)
> print(plot.ked.v, split = c(2, 1, 2, 1), more = F)
```



Q15 : *Can you see any evidence of the rock type map in the prediction map? If so, what?* *Jump to A15* •

It might be easier to answer the preceding question by showing the KED predictions and the rock types side-by-side:

```
> print(plot.ked, split = c(1, 1, 2, 1), more = T)
> print(map.rock, split = c(2, 1, 2, 1), more = F)
```



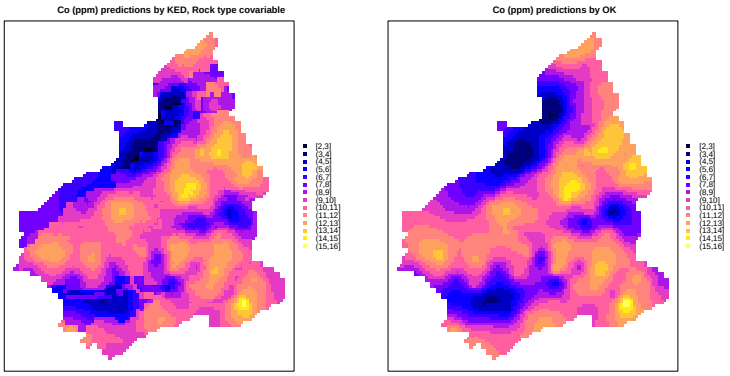
Task 12 : Compare the KED predictions to the OK predictions, on the same scale. •

First we compute the OK predictions for this grid:

```
> k.50 <- krige(Co ~ 1, loc = jura.cal, newdata = jura50,
+     model = vmf)

[using ordinary kriging]
```

Then we prepare a plot, using the same colour ramp and limits as for KED:

```
> plot.ok <- spplot(k.50, zcol = "var1.pred", contour = T,
+     pch = 15, by = seq(2, 16, by = 1), cuts = seq(2,
+         16, by = 1), col.regions = bpy.colors(64), key.space = "right",
+     main = "Co (ppm) predictions by OK")
> print(plot.ked, split = c(1, 1, 2, 1), more = T)
> print(plot.ok, split = c(2, 1, 2, 1), more = F)
```

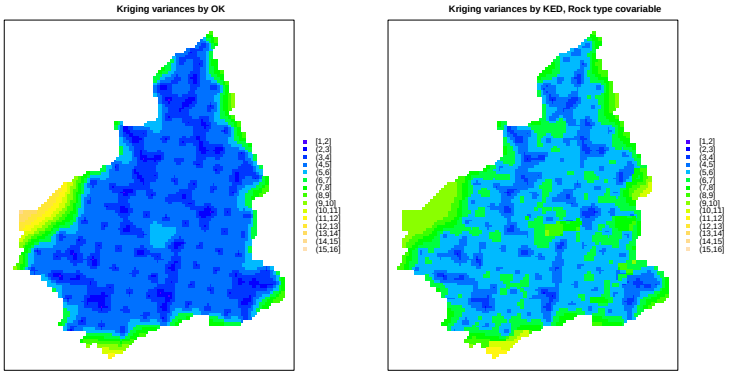


Q16 : *What are the principal differences between the **spatial pattern** of the two predictions?* *Jump to A16* •

Task 13 : Compare the kriging variances on the same scale. •

We already have the KED kriging variance map, so we just need to make a map on the same scale for OK:

```
> plot.ok.v <- spplot(k.50, zcol = "var1.var", contour = T,
+     pch = 15, cuts = seq(1, 16, by = 1), col.regions = topo.colors(64),
+     key.space = "right", main = "Kriging variances by OK")

> print(plot.ok.v, split = c(1, 1, 2, 1), more = T)
> print(plot.ked.v, split = c(2, 1, 2, 1), more = F)
```



Q17 : *Is there any difference in the **spatial pattern** of the two kriging variances? In the **values**? Explain why in both cases.* *Jump to A17* •

A useful way to visualize the differences is with a **difference map**.

Task 14 : Build a spatial data structure with the coördinates of the prediction grid, the differences in prediction and kriging variance between KED and OK, and the covariable (here, rock type). •

We use the `data.frame` method to build a data frame, and then the `coordinates` method to convert this to a spatial object:

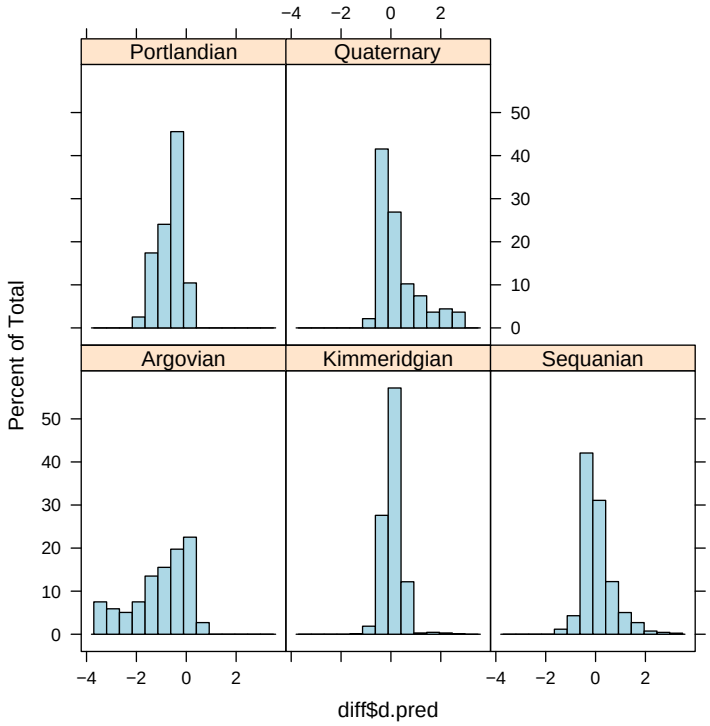
```
> diff <- data.frame(d.pred = kr.50$var1.pred - k.50$var1.pred,
+     d.var = kr.50$var1.var - k.50$var1.var, Rock = jura50$Rock)
> summary(diff)

     d.pred              d.var                 Rock
 Min.   :-3.4421   Min.   :-4.20   Argovian    :1185
 1st Qu.:-0.3973   1st Qu.: 1.03   Kimmeridgian:2036
 Median :-0.0939   Median : 1.14   Sequanian   :1628
 Mean   :-0.1658   Mean   : 1.10   Portlandian : 316
 3rd Qu.: 0.1922   3rd Qu.: 1.24   Quaternary  : 792
 Max.   : 3.2247   Max.   : 3.68

> coordinates(diff) <- coordinates(k.50)
```

We can visualize these differences both in feature and geographic space.

Task 15 : Display histograms of the differences in prediction for each rock type separately. •

We use the `histogram` method of the `lattice` package; like the `hist` method of the base package, this produces a histogram; but `histogram` has a **formula interface** to specify multiple plots. The so-called **conditioning variable**, which tells which categorical variable(s) to separate the plot by, is written to the right of a vertical bar `|`, and the variable to be plotted after the tilde `~`, to show this is not an x-y plot:

```
> print(histogram(~diff$d.pred | diff$Rock, col = "lightblue"))
```



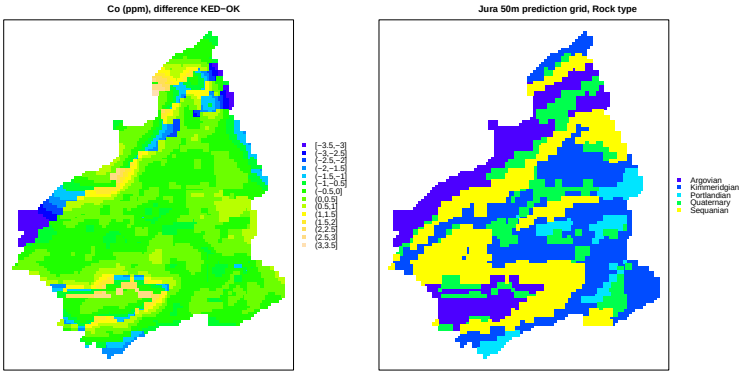
Pay close attention to the bin labels, especially the position of the zero point (i.e. zero residual). Note the differences are for KED – OK.

Q18 : *What are the differences in the distribution of the differences for the different rock types?* *Jump to A18* •

Of course we want to visualize the locations of the differences in geographic space.

Task 16 : Display a map of the differences, along with the map of rock types. •

```
> plot.diff <- spplot(diff, zcol = "d.pred", contour = T,
+     pch = 15, cuts = seq(-3.5, 3.5, by = 0.5), pretty = T,
+     col.regions = topo.colors(64), key.space = "right",
+     main = "Co (ppm), difference KED-OK")
> print(plot.diff, split = c(1, 1, 2, 1), more = T)
> print(map.rock, split = c(2, 1, 2, 1), more = F)
```



Q19 : *Where are the greatest negative and positive differences?* *Jump to A19* •

Finally, we examine the differences in kriging variance.

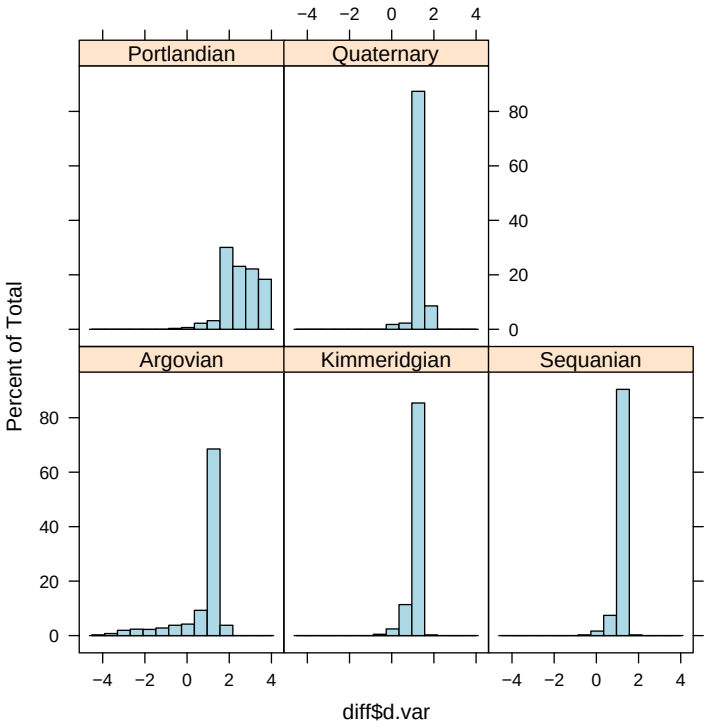
Task 17 : Summarize the differences in kriging variance. •

```
> summary(diff$d.pred)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.4400 -0.3970 -0.0939 -0.1660  0.1920  3.2200
```

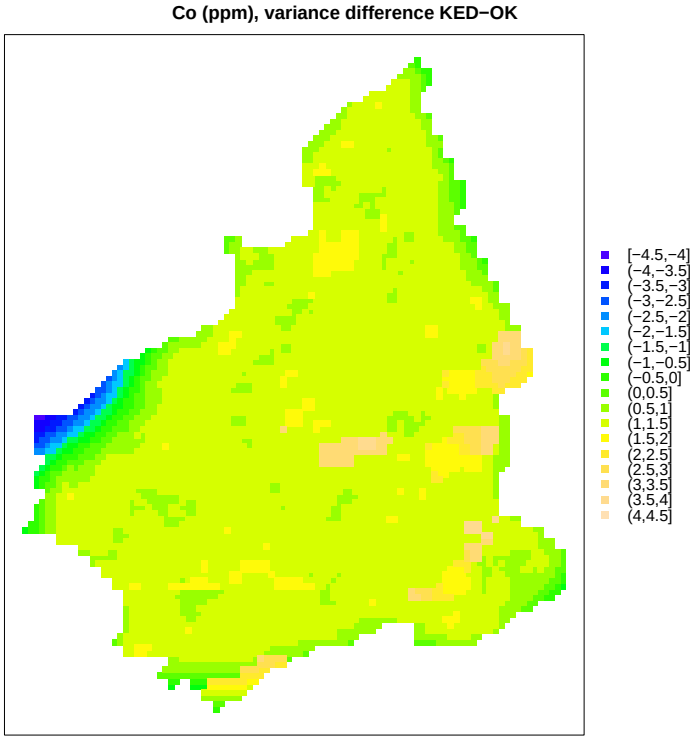
We can also visualize the histogram by rock type:

```
> print(histogram(~diff$d.var | diff$Rock, col = "lightblue"))
```



Task 18 : Display a map of the differences in kriging variance. •

```
> plot.diff.v <- spplot(diff, zcol = "d.var", contour = T,
+     pch = 15, cuts = seq(-4.5, 4.5, by = 0.5), pretty = T,
+     col.regions = topo.colors(64), key.space = "right",
+     main = "Co (ppm), variance difference KED-OK")
> print(plot.diff.v)
```



Q20 : *Where are the largest differences in kriging variance??* *Jump to A20* •

6.1 Answers for Kriging with external drift

A15 : *There are clear traces of the rock type, especially Argovian, which has substantially lower values than the others.* *Return to Q15* •

A16 : *The OK maps is quite smooth (continuous); the KED map shows discontinuities at boundaries between rock types (especially the Argovian with others).* *Return to Q16* •

A17 : *The spatial pattern is the same, because it depends only on sample point configuration, not on data or covariable (e.g. rock type) values. The variances are consistently lower for KED, because the total sill of the residual variogram is 35% lower than for the original variogram.* *Return to Q17* •

A18 : *The differences are minimal and symmetric in the Kimmeridgian; OK and KED agree closely. The Argovian has consistently negative differences, i.e OK > KED. This is because the class mean for Argovian is much lower than for the others, and KED takes this into account. The opposite is the case for the Quaternary.*

Return to Q18 •

A19 : *The biggest negative differences are associated with the Argovian rocks, positive with Quaternary.* *Return to Q19* •

A20 : *The KED variances tend to be lower overall (i.e. the difference KED – OK tends to be negative. The greatest improvement due to KED is in the W edge where there are no sample points* *Return to Q20* •

7 Validation

Both OK and KED can be validated with the 100 extra points in `jura.val` and cross-validated with the 259 points in the sample set `jura.cal`, just as we did for OK; these techniques will be introduced in Exercise 6 §2 and §3, respectively. You can then try them on the KED prediction to assess its quality.

8 Cleaning up

Task 19 : Remove the temporary variables from the workspace. •

```
> rm(map.rock, map.land)
> rm(plot.ok, plot.ked, plot.ok.v, plot.ked.v, plot.diff,
+     plot.diff.v)
```

9 Self-test

This section is a small self-test of how well you mastered this exercise. You should be able to complete the tasks and answer the questions with the knowledge you have gained from the exercise. Please **submit your answers (including graphical output) to the instructor** for grading and sample answers.

Task 1 : Compute and model the residual variogram using land use as the classifying factor. Plot the variogram and fitted model. Quantify the difference. •

Q1 : *How much does this differ from the ordinary variogram?* •

Q2 : *How much does this differ from the residual variogram using rock type as the classifying factor?* •

Q3 : *Explain why the residual variogram from land use differs from that using rock type.* •

Task 2 : Predict the Co concentration over the 50 m grid, using kriging with external drift (KED), with land use as the feature-space predictor. View the predictions and the kriging variances. •

Q4 : *Can you see any evidence of the land use map in the prediction map? If so, what?* •

Task 3 : Compare the KED predictions to the OK predictions, on the same scale. •

Q5 : *What are the principal differences between the* ***spatial pattern*** *of the two predictions? Is there a greater or lesser difference using land use as the feature-space predictor vs. using rock type? Why?* •

References

[1] P Goovaerts. *Geostatistics for natural resources evaluation*. Applied Geostatistics. Oxford University Press, New York; Oxford, 1997. 11

Index of R Concepts

<- operator, 3

anova, 6

boxplot, 2

ceiling, 8
coordinates (package:sp), 16

data.frame, 16

fit.variogram (package:gstat), 9

gstat package, 2, 7, 9, 11

hist, 17
histogram (package:lattice), 17

jura dataset, 11

krige (package:gstat), 14
krige package, 14

lattice package, 17
lm, 3, 8
load, 2, 11

setwd, 2
sp package, 2
spplot (package:sp), 12
summary, 3
summary.lm, 3

table, 12

variogram (package:gstat), 7
vgm (package:gstat), 8