2

Regression-kriging

As we saw in the previous chapter, there are many geostatistical techniques that can be used to map environmental variables. In reality, we always try to go for the most flexible, most comprehensive and the most robust technique (preferably implemented in a software with an user-friendly GUI). In fact, many (geo)statisticians believe that there is only one Best Linear Unbiased Prediction (**BLUP**) model for spatial data, from which all other (linear) techniques can be derived (Gotway and Stroup, 1997; Stein, 1999; Christensen, 2001). As we will see in this chapter, one such generic mapping technique is regression-kriging. All other techniques mentioned previously — ordinary kriging, environmental correlation, averaging of values per polygons or inverse distance interpolation — can be seen as special cases of RK.

2.1 The Best Linear Unbiased Predictor of spatial data

Matheron (1969) proposed that a value of a target variable at some location can be modeled as a sum of the deterministic and stochastic components:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon'(\mathbf{s}) + \varepsilon'' \tag{2.1.1}$$

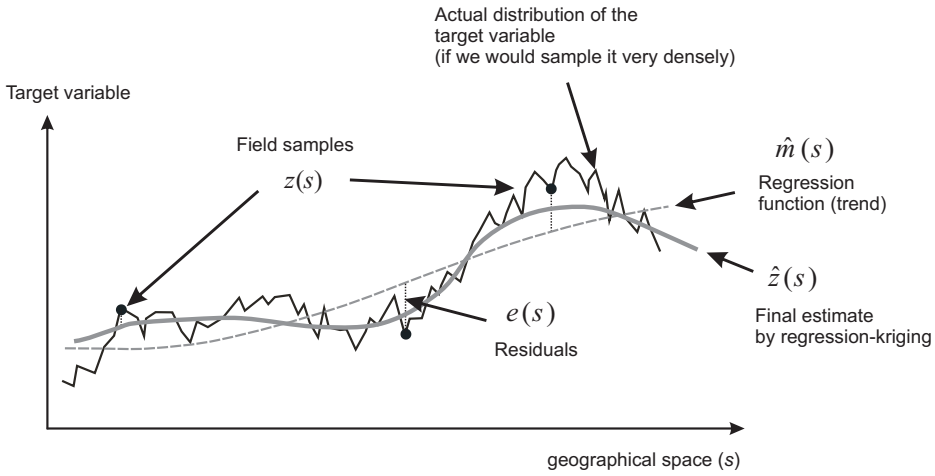

Fig. 2.1: A schematic example of the regression-kriging concept shown using a cross-section.

which he termed the **universal model of spatial variation**. We have seen in the previous sections (§1.3.1 and §1.3.2) that both deterministic and stochastic components of spatial variation can be modeled separately. By combining the two approaches, we obtain:

$$\begin{aligned}\hat{z}(\mathbf{s}_0) &= \hat{m}(\mathbf{s}_0) + \hat{e}(\mathbf{s}_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\mathbf{s}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\mathbf{s}_i)\end{aligned} \tag{2.1.2}$$

where $\hat{m}(\mathbf{s}_0)$ is the fitted deterministic part, $\hat{e}(\mathbf{s}_0)$ is the interpolated residual, $\hat{\beta}_k$ are estimated deterministic model coefficients ($\hat{\beta}_0$ is the estimated intercept), λ_i are kriging weights determined by the spatial dependence structure of the residual and where $e(\mathbf{s}_i)$ is the residual at location $\mathbf{s}_i$. The regression coefficients $\hat{\beta}_k$ can be estimated from the sample by some fitting method, e.g. ordinary least squares (OLS) or, optimally, using **Generalized Least Squares** (Cressie, 1993, p.166):

$$\hat{\beta}_{\mathtt{GLS}} = \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{z} \tag{2.1.3}$$

where $\hat{\beta}_{\mathtt{GLS}}$ is the vector of estimated regression coefficients, $\mathbf{C}$ is the covariance matrix of the residuals, $\mathbf{q}$ is a matrix of predictors at the sampling locations and $\mathbf{z}$ is the vector of measured values of the target variable. The GLS estimation of regression coefficients is, in fact, a special case of geographically weighted regression (compare with Eq.1.3.20). In this case, the weights are determined objectively to account for the spatial auto-correlation between the residuals.

Once the deterministic part of variation has been estimated, the residual can be interpolated with kriging and added to the estimated trend (Fig. 2.1). Estimation of the residuals and their variogram model is an iterative process: first the deterministic part of variation is estimated using ordinary least squares (OLS), then the covariance function of the residuals is used to obtain the GLS coefficients. Next, these are used to re-compute the residuals, from which an updated covariance function is computed, and so on (Schabenberger and Gotway, 2004, p.286). Although this is recommended as the proper procedure by many geostatisticians, Kitanidis (1994) showed that use of the covariance function derived from the OLS residuals (i.e. a single iteration) is often satisfactory, because it is not different enough from the function derived after several iterations; i.e. it does not affect the final predictions much. Minasny and McBratney (2007) reported similar results: it is often more important to use more useful and higher quality data than to use more sophisticated statistical methods. In some situations[1] however, the model needs to be fitted using the most sophisticated technique to avoid making biased predictions.

In matrix notation, regression-kriging is commonly written as (Christensen, 2001, p.277):

$$\hat{z}_{\mathtt{RK}}(\mathbf{s}_0) = \mathbf{q}_{\mathbf{0}}^{\mathbf{T}} \cdot \hat{\beta}_{\mathtt{GLS}} + \lambda_{\mathbf{0}}^{\mathbf{T}} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}_{\mathtt{GLS}}) \tag{2.1.4}$$

where $\hat{z}(\mathbf{s}_0)$ is the predicted value at location $\mathbf{s}_0$, $\mathbf{q_0}$ is the vector of $p+1$ predictors and $\lambda_{\mathbf{0}}$ is the vector of n kriging weights used to interpolate the residuals. The model in Eq.(2.1.4) is considered to be the Best Linear Predictor of spatial data (Christensen, 2001; Schabenberger and Gotway, 2004). It has a prediction variance that reflects the position of new locations (extrapolation effect) in both geographical and feature space:

$$\hat{\sigma}^2_{\mathtt{RK}}(\mathbf{s}_0) = (C_0 + C_1) - \mathbf{c}_{\mathbf{0}}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0} + \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right)^{\mathbf{T}} \cdot \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right) \tag{2.1.5}$$

where $C_0 + C_1$ is the sill variation and $\mathbf{c}_0$ is the vector of covariances of residuals at the unvisited location.

[1] For example: if the points are extremely clustered, and/or if the sample is $\ll$100, and/or if the measurements are noisy or obtained using non-standard techniques.

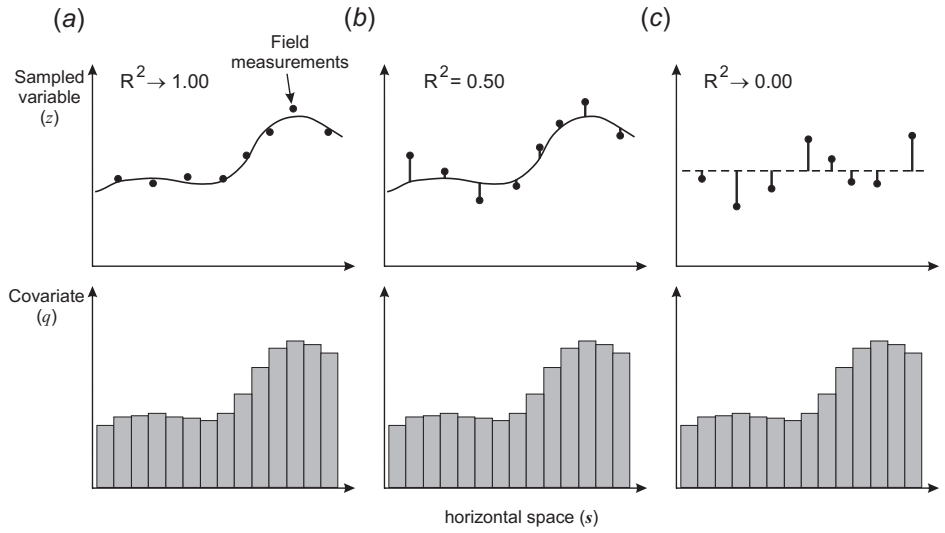

Fig. 2.2: Whether we will use pure regression model, pure kriging or hybrid regression-kriging is basically determined by R-square: (a) if R-square is high, then the residuals will be infinitively small; (c) if R-square is insignificant, then we will probably finish with using ordinary kriging; (b) in most cases, we will use a combination of regression and kriging.

If the residuals show no spatial auto-correlation (pure nugget effect), the regression-kriging (Eq.2.1.4) converges to pure multiple linear regression (Eq.1.3.14) because the covariance matrix (**C**) becomes identity matrix:

$$\mathbf{C} = \begin{bmatrix} C_0 + C_1 & \cdots & 0 \\ \vdots & C_0 + C_1 & 0 \\ 0 & 0 & C_0 + C_1 \end{bmatrix} = (C_0 + C_1) \cdot \mathbf{I} \tag{2.1.6}$$

so the kriging weights (Eq.1.3.4) at any location predict the mean residual i.e. 0 value. Similarly, the regression-kriging variance (Eq.2.1.5) reduces to the multiple linear regression variance (Eq.1.3.16):

$$\hat{\sigma}^2_{\mathrm{RK}}(\mathbf{s}_0) = (C_0 + C_1) - 0 + \mathbf{q}_0^{\mathrm{T}} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \frac{1}{(C_0 + C_1)} \cdot \mathbf{q} \right)^{-1} \cdot \mathbf{q}_0$$

$$\hat{\sigma}^2_{\mathrm{RK}}(\mathbf{s}_0) = (C_0 + C_1) + (C_0 + C_1) \cdot \mathbf{q}_0^{\mathrm{T}} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{q} \right)^{-1} \cdot \mathbf{q}_0$$

and since $(C_0 + C_1) = C(0) = MSE$, the RK variance reduces to the MLR variance:

$$\hat{\sigma}^2_{\mathrm{RK}}(\mathbf{s}_0) = \hat{\sigma}^2_{\mathrm{OLS}}(\mathbf{s}_0) = MSE \cdot \left[1 + \mathbf{q}_0^{\mathrm{T}} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{q} \right)^{-1} \cdot \mathbf{q}_0 \right] \tag{2.1.7}$$

Likewise, if the target variable shows no correlation with the auxiliary predictors, the regression-kriging model reduces to ordinary kriging model because the deterministic part equals the (global) mean value (Fig. 2.2c, Eq.1.3.25).

The formulas above show that, depending on the strength of the correlation, RK might turn into pure kriging — if predictors are uncorrelated with the target variable — or pure regression — if there is significant correlation and the residuals show pure nugget variogram (Fig. 2.2). Hence, pure kriging and pure regression should be considered as only special cases of regression-kriging (Hengl et al., 2004a, 2007a).

2.1.1 Mathematical derivation of BLUP

Understanding how a prediction model is derived becomes important once we start getting strange results or poor cross-validation scores. Each model is based on some assumptions that need to be respected and taken into account during the final interpretation of results. A detailed derivation of the BLUP for spatial data can be followed in several standard books on geostatistics (Stein, 1999; Christensen, 2001); one of the first complete derivations is given by Goldberger (1962). Here is a somewhat shorter explanation of how BLUP is derived, and what the implications of various mathematical assumptions are.

All flavors of linear statistical predictors share the same objective of minimizing the estimation error variance $\hat{\sigma}^2_E(\mathbf{s}_0)$ under the constraint of unbiasedness (Goovaerts, 1997). In mathematical terms, the estimation error:

$$\hat{\sigma}^2(\mathbf{s}_0) = E\left\{\left(\hat{z}(\mathbf{s}_0) - z(\mathbf{s}_0)\right) \cdot \left(\hat{z}(\mathbf{s}_0) - z(\mathbf{s}_0)\right)^T\right\} \tag{2.1.8}$$

is minimized under the (unbiasedness) constraint that:

$$E\left\{\hat{z}(\mathbf{s}_0) - z(\mathbf{s}_0)\right\} = 0 \tag{2.1.9}$$

Assuming the universal model of spatial variation, we can define a generalized linear regression model (Goldberger, 1962):

$$z(\mathbf{s}) = \mathbf{q}^{\mathrm{T}} \cdot \beta + \varepsilon(\mathbf{s}) \tag{2.1.10}$$
$$E\left\{\varepsilon(\mathbf{s})\right\} = 0 \tag{2.1.11}$$
$$E\left\{\varepsilon \cdot \varepsilon^{\mathrm{T}}(\mathbf{s})\right\} = \mathbf{C} \tag{2.1.12}$$

where ε is the residual variation, and $\mathbf{C}$ is the $n \times n$ positive-definite variance-covariance matrix of residuals. This model can be read as follows: (1) the information signal is a function of deterministic and residual parts; (2) the best estimate of the residuals is 0; (3) the best estimate of the correlation structure of residuals is the variance-covariance matrix.

Now that we have defined the statistical model and the minimization criteria, we can derive the best linear unbiased prediction of the target variable:

$$\hat{z}(\mathbf{s}_0) = \hat{\delta}_{\mathbf{0}}^{\mathrm{T}} \cdot \mathbf{z} \tag{2.1.13}$$

Assuming that we use the model shown in Eq.(2.1.10), and assuming that the objective is to minimize the estimation error $\hat{\sigma}^2_E(\mathbf{s}_0)$, it can be shown[2] that BLUP parameters can be obtained by solving the following system:

$$\begin{bmatrix} \mathbf{C} & \mathbf{q} \\ \mathbf{q}^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \delta \\ \phi \end{bmatrix} = \begin{bmatrix} \mathbf{c_0} \\ \mathbf{q_0} \end{bmatrix} \tag{2.1.14}$$

where $\mathbf{c_0}$ is the vector of $n \times 1$ covariances at a new location, $\mathbf{q_0}$ is the vector of $p \times 1$ predictors at a new location, and ϕ is a vector of Lagrange multipliers. It can be further shown that, by solving the Eq.(2.1.14), we get the following:

[2] The actual derivation of formulas is not presented. Readers are advised to obtain the paper by Goldberger (1962).

$$\begin{aligned}\hat{z}(\mathbf{s}_0) &= \mathbf{q}_0^{\mathrm{T}} \cdot \hat{\beta} + \hat{\lambda}_0^{\mathrm{T}} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}) \\ \hat{\beta} &= \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{z} \\ \hat{\lambda}_0 &= \mathbf{C}^{-1} \cdot \mathbf{c}_0\end{aligned} \tag{2.1.15}$$

which is the prediction model explained in the previous section.

Under the assumption of the **first order stationarity** i.e. constant trend:

$$E\{z(\mathbf{s})\} = \mu \qquad \forall \mathbf{s} \in \mathbb{A} \tag{2.1.16}$$

the Eq.(2.1.15) modifies to (Schabenberger and Gotway, 2004, p.268):

$$\begin{aligned}\hat{z}(\mathbf{s}_0) &= \mu + \hat{\lambda}_0^{\mathrm{T}} \cdot (\mathbf{z} - \mu) \\ \hat{\lambda}_0 &= \mathbf{C}^{-1} \cdot \mathbf{c}_0\end{aligned}$$

i.e. to ordinary kriging (§1.3.1). If we assume that the deterministic part of variation is not constant, then we need to consider obtaining a number of covariates (**q**) that can be used to model the varying mean value.

Another important issue you need to know about the model in Eq.(2.1.15) is that, in order to solve the residual part of variation, we need to know covariances at new locations:

$$C\left(e(\mathbf{s}_0), e(\mathbf{s}_i)\right) = E\left[\{e(\mathbf{s}_0) - \mu\} \cdot \{e(\mathbf{s}_i) - \mu\}\right] \tag{2.1.17}$$

which would require that we know the values of the target variable at a new location ($e(\mathbf{s}_0)$), which we of course do not know. Instead, we can use the existing sampled values ($e(\mathbf{s}_i) = z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i)$) to model the covariance structure using a pre-defined mathematical model (e.g. Eq.1.3.8). If we assume that the covariance model is the same (constant) in the whole area of interest, then the covariance is dependent only on the separation vector **h**:

$$C\left(e(\mathbf{s}_0), e(\mathbf{s}_i)\right) = C(\mathbf{h}) \tag{2.1.18}$$

which is known as the assumption of **second order stationarity**; and which means that we can use the same model to predict values anywhere in the area of interest (global estimation). If this assumption is not correct, we would need to estimate covariance models locally. This is often not so trivial because we need to have a lot of points (see further §2.2), so the assumption of second order stationarity is very popular among geostatisticians. Finally, you need to also be aware that the residuals in Eq.(2.1.10) are expected to be normally distributed around the regression line and homoscedastic[3], as with any linear regression model (Kutner et al., 2004). If this is not the case, then the target variable needs to be transformed until these conditions are met.

The first and second order stationarity, and normality of residuals/target variables are rarely tested in real case studies. In the case of regression-kriging (see further §2.1), the target variable does not have to be stationary but its residuals do, hence we do not have to test this property with the original variable. In the case of regression-kriging in a moving window, we do not have to test neither first nor second order stationarity. Furthermore, if the variable is non-normal, then we can use some sort of GLM to fit the model. If this is successful, the residuals will typically be normal around the regression line in the transformed space, and this will allow us to proceed with kriging. The predicted values can finally be back-transformed to the original scale using the inverse of the link function.

[3] Meaning symmetrically distributed around the feature space and the regression line.

The lesson learned is that each statistical spatial predictor comes with: (a) a conceptual model that explains the general relationship (e.g. Eq.2.1.10); (b) model-associated assumptions (e.g. zero mean estimation error, first or second order stationarity, normality of the residuals); (c) actual prediction formulas (Eq.2.1.15); and (d) a set of proofs that, under given assumptions, a prediction model is the BLUP. Ignoring the important model assumptions can lead to poor predictions, even though the output maps might appear to be visually fine.

2.1.2 Selecting the right spatial prediction technique

Knowing that the most of the linear spatial prediction models are more or less connected, we can start by testing the most generic technique, and then finish by using the most suitable technique for our own case study. Pebesma (2004, p.689), for example, implemented such a nested structure in his design of the `gstat` package. An user can switch between one and another technique by following a simple decision tree shown in Fig. 2.3.

First, we need to check if the deterministic model is defined already, if it has not been, we can try to correlate the sampled variables with environmental factors. If the environmental factors are significantly correlated, we can fit a multiple linear regression model (Eq.1.3.14) and then analyze the residuals for spatial autocorrelation. If the residuals show no spatial autocorrelation (pure nugget effect), we proceed with OLS estimation of the regression coefficients. Otherwise, if the residuals show spatial auto-correlation, we can run regression-kriging. If the data shows no correlation with environmental factors, then we can still analyze the variogram of the target variable. This time, we might also consider modeling the anisotropy. If we can fit a variogram different from pure nugget effect, then we can run ordinary kriging. Otherwise, if we can only fit a linear variogram, then we might just use some mechanical interpolator such as the inverse distance interpolation.

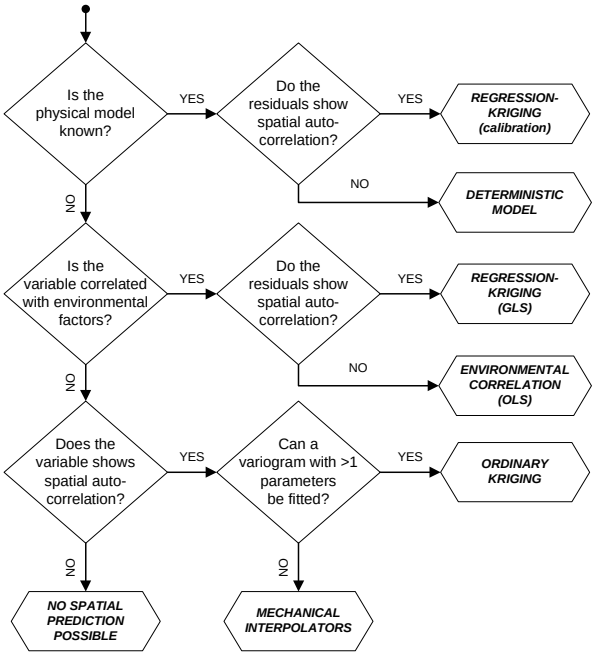

Fig. 2.3: A general decision tree for selecting the suitable spatial prediction model based on the results of model estimation. Similar decision tree is implemented in the `gstat` package.

If the variogram of the target variable shows no spatial auto-correlation, and no correlation with environmental factors, this practically means that the only statistically valid prediction model is to estimate a global mean for the whole area. Although this might frustrate you because it would lead to a nonsense map where each pixel shows the same value, you should be aware that even this is informative[4].

How does the selection of the spatial prediction model works in practice? In the `gstat` package, a user can easily switch from one to other prediction model by changing the arguments in the generic `krige` function in R (Fig. 1.13; see further §3.2). For example, if the name of the input field samples is `meuse` and the prediction locations (grid) is defined by `meuse.grid`, we can run the inverse distance interpolation (§1.2.1) by specifying (Pebesma, 2004):

```
> library(gstat)
> data(meuse)
> coordinates(meuse) <- ~ x+y
> data(meuse.grid)
> coordinates(meuse.grid) <- ~ x+y
```

[4] Sometimes an information that we are completely uncertain about a feature is better than a colorful but completely unreliable map.

```
> gridded(meuse.grid) <- TRUE
> zinc.id <- krige(zinc ~ 1, data=meuse, newdata=meuse.grid)

  [inverse distance weighted interpolation]
```

where zinc is the sampled environmental variable (vector) and zinc.id is the resulting raster map (shown in Fig. 1.13). Instead of using inverse distance interpolation we might also try to fit the values using the coordinates and a 2nd order polynomial model:

```
> zinc.ts <- krige(zinc ~ x+y+x*y+x*x+y*y, data=meuse, newdata=meuse.grid)

  [ordinary or weighted least squares prediction]
```

which can be converted to the moving surface fitting by adding a search window:

```
> zinc.mv <- krige(zinc ~ x+y+x*y+x*x+y*y, data=meuse, newdata=meuse.grid, nmax=20)

  [ordinary or weighted least squares prediction]
```

If we add a variogram model, then gstat will instead of running inverse distance interpolation run ordinary kriging (§1.3.1):

```
> zinc.ok <- krige(log1p(zinc) ~ 1, data=meuse, newdata=meuse.grid,
+            model=vgm(psill=0.714, "Exp", range=449, nugget=0))

  [using ordinary kriging]
```

where vgm(psill=0.714, "Exp", range=449, nugget=0) is the Exponential variogram model with a sill parameter of 0.714, range parameter of 449 m and the nugget parameter of 0 (the target variable was log-transformed). Likewise, if there were environmental factors significantly correlated with the target variable, we could run OLS regression (§1.3.2) by omitting the variogram model:

```
> zinc.ec <- krige(log1p(zinc) ~ dist+ahn, data=meuse, newdata=meuse.grid)

  [ordinary or weighted least squares prediction]
```

where dist and ahn are the environmental factor used as predictors (raster maps), which are available as separate layers within the spatial layer[5] meuse.grid. If the residuals do show spatial auto-correlation, then we can switch to universal kriging (Eq.2.1.4) by adding the variogram:

```
> zinc.rk <- krige(log1p(zinc) ~ dist+ahn, data=meuse, newdata=meuse.grid,
+            model=vgm(psill=0.151, "Exp", range=374, nugget=0.055))

  [using universal kriging]
```

If the model between the environmental factors and our target variable is deterministic, then we can use the point samples to calibrate our predictions. The R command would then look something like this:

```
> zinc.rkc <- krige(zinc ~ zinc.df, data=meuse, newdata=meuse.grid,
+             model=vgm(psill=3, "Exp", range=500, nugget=0))

  [using universal kriging]
```

where zinc.df are the values of the target variable estimated using a deterministic function.

In gstat, a user can also easily switch from estimation to simulations (§2.4) by adding to the command above an additional argument: nsim=1. This will generate Sequential Gaussian Simulations using the same prediction model. Multiple simulations can be generated by increasing the number set for this argument. In addition, a user can switch from block predictions by adding argument block=100; and from global estimation of weights by adding a search radius or maximum number of pairs, e.g. radius=1000 or nmax=60.

By using the automap[6] package one needs to specify even less arguments. For example, the command:

[5] In R a SpatialGridDataframe object.

[6] http://cran.r-project.org/web/packages/automap/

```
> zinc.rk <- autoKrige(log1p(zinc) ~ dist, data=meuse, newdata=meuse.grid)

  [using universal kriging]
```

will do much of the standard geostatistical analysis without any intervention from the user: it will filter the duplicate points where needed, estimate the residuals, then fit the variogram for the residuals, and generate the predictions at new locations. The results can be plotted in a single page in a form of a report. Such generic commands can significantly speed up data processing, and make it easier for a non-geostatistician to generate maps (see further section 2.10.3).

In the intamap package[7], one needs to set even less parameters to generate predictions from a variety of methods:

```
> meuse$value <- log(meuse$zinc)
> output <- interpolate(data=meuse, newdata=meuse.grid)

  R 2009-11-11 17:09:14 interpolating 155 observations, 3103 prediction locations
  [Time models loaded...]
  [1] "estimated time for  copula 133.479866956255"
  Checking object ... OK
```

which gives the (presumably) best interpolation method[8] for the current problem (`value` column), given the time available set with `maximumTime`.

A more systematic strategy to select the right spatial prediction technique is to use objective criteria of mapping success (i.e. *a posteriori* criteria). From the application point of view, it can be said that there are (only) five relevant criteria to evaluate various spatial predictors (see also §1.4):

(1.) the **overall mapping accuracy**, e.g. standardized RMSE at control points — the amount of variation explained by the predictor expressed in %;

(2.) the **bias**, e.g. mean error — the accuracy of estimating the central population parameters;

(3.) the **model robustness**, also known as *model sensitivity* — in how many situations would the algorithm completely fail / how much artifacts does it produces?;

(4.) the **model reliability** — how good is the model in estimating the prediction error (how accurate is the prediction variance considering the true mapping accuracy)?;

(5.) the **computational burden** — the time needed to complete predictions;

From this five, you could derive a single composite measure that would then allow you to select '*the optimal*' predictor for any given data set, but this is not trivial! Hsing-Cheng and Chun-Shu (2007) suggest a framework to select the best predictor in an automated way, but this work would need to be much extended. In many cases we simply finish using some **naïve predictor** — that is predictor that we know has a statistically more optimal alternative[9], but this alternative is simply not practical.

The intamap decision tree, shown in Fig. 2.4, is an example of how the selection of the method can be automated to account for (1) anisotropy, (2) specified observation errors, and (3) extreme values. This is a specific application primarily developed to interpolate the radioactivity measurements from the European radiological data exchange platform, a network of around 4000 sensors. Because the radioactivity measurements can often carry local extreme values, robust techniques need to be used to account for such effects. For example, **Copula kriging**[10] methods can generate more accurate maps if extreme values are also present in the observations. The problem of using methods such as Copula kriging, however, is that they can often take even few hours to generate maps even for smaller areas. To minimize the risk of running into endless computing, the authors of the intamap decision tree have decided to select the prediction algorithm based on

[7] `http://cran.r-project.org/web/packages/intamap/`

[8] intamap automatically chooses between: (1) kriging, (2) copula methods, (3) inverse distance interpolation, projected spatial gaussian process methods in the psgp package, (4) transGaussian kriging or yamamoto interpolation.

[9] For example, instead of using the REML approach to variogram modeling, we could simply fit a variogram using weighted least squares (see §1.3.1), and ignore all consequences (Minasny and McBratney, 2007).

[10] Copula kriging is a sophistication of ordinary kriging; an iterative technique that splits the original data set and then re-estimates the model parameters with maximization of the corresponding likelihood function (Bárdossy and Li, 2008).

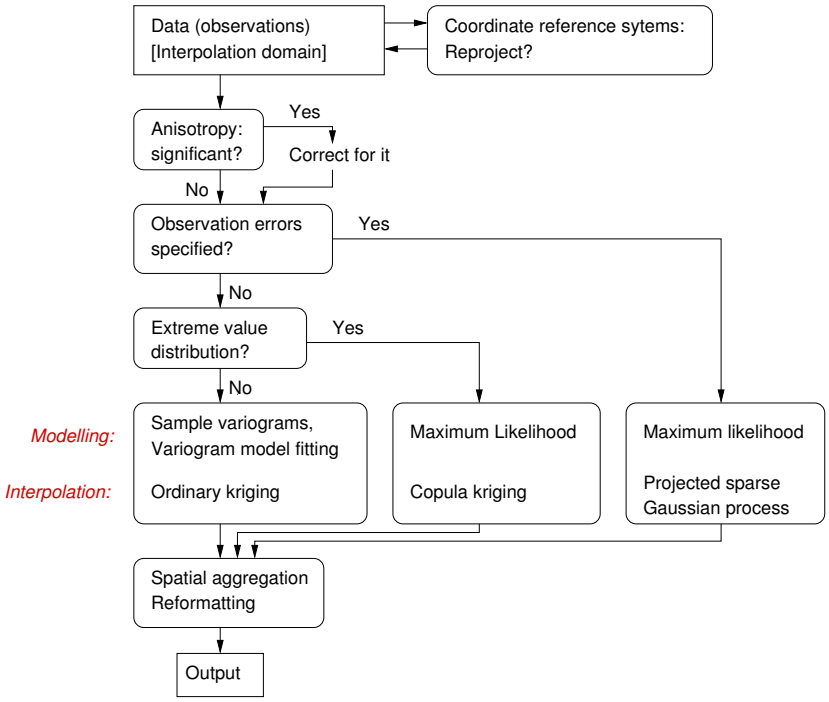

Fig. 2.4: Decision tree used in the intamap interpolation service for automated mapping. After Pebesma et al. (2009).

the computational time. Hence the system first estimates the approximate time needed to run the prediction using the most sophisticated technique; if this is above the threshold time, the system will switch to a more naïve method (Pebesma et al., 2009). As a rule of thumb, the authors of intamap suggest 30 seconds as the threshold time to accept automated generation of a map via a web-service.

2.1.3 The Best Combined Spatial Predictor

Assuming that a series of prediction techniques are mutually independent[11], predictions can be generated as a weighted average from multiple predictions i.e. by generating the Best Combined Spatial Prediction (**BCSP**):

$$\hat{z}_{\text{BCSP}}(\mathbf{s_0}) = \frac{\hat{z}_{\text{SP1}}(\mathbf{s_0}) \cdot \frac{1}{\hat{\sigma}_{\text{SP1}}(\mathbf{s_0})} + \hat{z}_{\text{SP2}}(\mathbf{s_0}) \cdot \frac{1}{\hat{\sigma}_{\text{SP2}}(\mathbf{s_0})} + \ldots + \hat{z}_{\text{SPj}}(\mathbf{s_0}) \cdot \frac{1}{\hat{\sigma}_{\text{SPj}}(\mathbf{s_0})}}{\sum_{j=1}^{p} \frac{1}{\hat{\sigma}_{\text{SPj}}(\mathbf{s_0})}} \tag{2.1.19}$$

where $\sigma_{\text{SPj}}(\mathbf{s_0})$ is the prediction error estimated by the model (prediction variance), and p is the number of predictors. For example, we can generate a combined prediction using OK and e.g. GLM-regression and then sum-up the two maps (Fig. 2.5). The predictions will in some parts of the study are look more as OK, in others more as GLM, which actually depicts extrapolation areas of both methods. This map is very similar to predictions produced using regression-kriging (see further Fig. 5.9); in fact, one could probably mathematically prove that under ideal conditions (absolute stationarity of residuals; no spatial clustering; perfect linear relationship), BCSP predictions would equal the regression-kriging predictions. In general, the map in the middle of Fig. 2.5 looks more as the GLM-regression map because this map is about 2–3 times more precise than the OK map. It is important to emphasize that, in order to combine various predictors, we do need to have an estimate of the prediction uncertainty, otherwise we are not able to assign the weights (see further §7.5). In principle, linear combination of statistical techniques using the Eq.(2.1.19) above should be avoided if a theoretical basis exists that incorporates such combination.

[11] If they do not use the same model parameters; if they treat different parts of spatial variation etc.

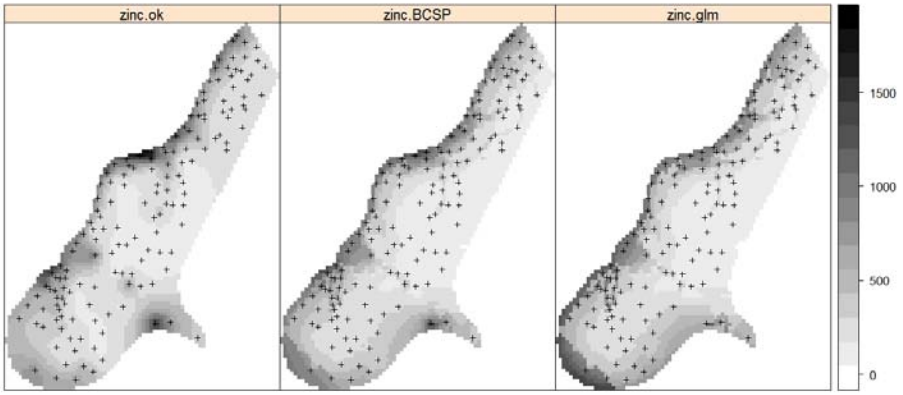

Fig. 2.5: Best Combined Spatial Predictor as weighted average of ordinary kriging (`zinc.ok`) and GLM regression (`zinc.glm`).

In the example above (GLM+OK), we assume that the predictions/prediction errors are independent, and they are probably not. In addition, a statistical theory exists that supports a combination of regression and kriging (see previously §2.1.1), so there is no need to run predictions separately and derive an unrealistic measure of model error. The BCSP can be only interesting for situations where there are indeed several objective predictors possible, where no theory exists that reflects their combination, and/or where fitting of individual models is faster and less troublesome than fitting of a hybrid model. For example, ordinary kriging can be speed-up by limiting the search radius, predictions using GLMs is also relatively inexpensive. External trend kriging using a GLM in geoRglm package might well be the statistically most robust technique you could possibly use, but it can also be beyond the computational power of your PC.

The combined prediction error of a BCSP can be estimated as the smallest prediction error achieved by any of the prediction models:

$$\hat{\sigma}_{\texttt{BCSP}}(\mathbf{s_0}) = \min\left\{\hat{\sigma}_{\texttt{SP1}}(\mathbf{s_0}), \ldots, \hat{\sigma}_{\texttt{SPj}}(\mathbf{s_0})\right\} \tag{2.1.20}$$

which is really an *ad hoc* formula and should be used only to visualize and depict problematic areas (highest prediction error).

2.1.4 Universal kriging, kriging with external drift

The geostatistical literature uses many different terms for what are essentially the same or at least very similar techniques. This confuses the users and distracts them from using the right technique for their mapping projects. In this section, we will show that both universal kriging, kriging with external drift and regression-kriging are basically the same technique. Matheron (1969) originally termed the technique *Le krigeage universel*, however, the technique was intended as a generalized case of kriging where the trend is modeled as a function of coordinates. Thus, many authors (Deutsch and Journel, 1998; Wackernagel, 2003; Papritz and Stein, 1999) reserve the term *Universal Kriging* (UK) for the case when only the coordinates are used as predictors. If the deterministic part of variation (*drift*) is defined externally as a linear function of some explanatory variables, rather than the coordinates, the term *Kriging with External Drift* (KED) is preferred (Wackernagel, 2003; Chiles and Delfiner, 1999). In the case of UK or KED, the predictions are made as with kriging, with the difference that the covariance matrix of residuals is extended with the auxiliary predictors $q_k(\mathbf{s}_i)$'s (Webster and Oliver, 2001, p.183). However, the drift and residuals can also be estimated separately and then summed. This procedure was suggested by Ahmed and de Marsily (1987); Odeh et al. (1995) later named it *Regression-kriging*, while Goovaerts (1997, §5.4) uses the term *Kriging with a trend model* to refer to a family of predictors, and refers to RK as *Simple kriging with varying local means*. Although equivalent, KED and RK differ in the computational steps used.

Let us zoom into the two variants of regression-kriging. In the case of KED, predictions at new locations are made by:

$$\hat{z}_{\text{KED}}(\mathbf{s}_0) = \sum_{i=1}^{n} w_i^{\text{KED}}(\mathbf{s}_0) \cdot z(\mathbf{s}_i) \tag{2.1.21}$$

for

$$\sum_{i=1}^{n} w_i^{\text{KED}}(\mathbf{s}_0) \cdot q_k(\mathbf{s}_i) = q_k(\mathbf{s}_0); \qquad k = 1, ..., p \tag{2.1.22}$$

or in matrix notation:

$$\hat{z}_{\text{KED}}(\mathbf{s}_0) = \delta_{\mathbf{0}}^{\mathbf{T}} \cdot \mathbf{z} \tag{2.1.23}$$

where z is the target variable, q_k's are the predictor variables i.e. values at a new location ($\mathbf{s}_0$), $\delta_{\mathbf{0}}$ is the vector of KED weights (w_i^{KED}), p is the number of predictors and $\mathbf{z}$ is the vector of n observations at primary locations. The KED weights are solved using the extended matrices:

$$\begin{aligned} \lambda_{\mathbf{0}}^{\text{KED}} &= \left\{ w_1^{\text{KED}}(\mathbf{s}_0), ..., w_n^{\text{KED}}(\mathbf{s}_0), \varphi_0(\mathbf{s}_0), ..., \varphi_p(\mathbf{s}_0) \right\}^{\mathbf{T}} \\ &= \mathbf{C}^{\text{KED}-1} \cdot \mathbf{c}_{\mathbf{0}}^{\text{KED}} \end{aligned} \tag{2.1.24}$$

where $\lambda_{\mathbf{0}}^{\text{KED}}$ is the vector of solved weights, φ_p are the Lagrange multipliers, $\mathbf{C}^{\text{KED}}$ is the extended covariance matrix of residuals and $\mathbf{c}_{\mathbf{0}}^{\text{KED}}$ is the extended vector of covariances at a new location.

In the case of KED, the extended covariance matrix of residuals looks like this (Webster and Oliver, 2001, p.183):

$$\mathbf{C}^{\text{KED}} = \begin{bmatrix} C(\mathbf{s}_1, \mathbf{s}_1) & \cdots & C(\mathbf{s}_1, \mathbf{s}_n) & 1 & q_1(\mathbf{s}_1) & \cdots & q_p(\mathbf{s}_1) \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ C(\mathbf{s}_n, \mathbf{s}_1) & \cdots & C(\mathbf{s}_n, \mathbf{s}_n) & 1 & q_1(\mathbf{s}_n) & \cdots & q_p(\mathbf{s}_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ q_1(\mathbf{s}_1) & \cdots & q_1(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & 0 & \vdots & & \vdots \\ q_p(\mathbf{s}_1) & \cdots & q_p(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{2.1.25}$$

and $\mathbf{c}_{\mathbf{0}}^{\text{KED}}$ like this:

$$\mathbf{c}_{\mathbf{0}}^{\text{KED}} = \left\{ C(\mathbf{s}_0, \mathbf{s}_1), ..., C(\mathbf{s}_0, \mathbf{s}_n), q_0(\mathbf{s}_0), q_1(\mathbf{s}_0), ..., q_p(\mathbf{s}_0) \right\}^{\mathbf{T}}; \quad q_0(\mathbf{s}_0) = 1 \tag{2.1.26}$$

Hence, KED looks exactly as ordinary kriging (Eq.1.3.2), except the covariance matrix and vector are extended with values of auxiliary predictors.

In the case of RK, the predictions are made separately for the drift and residuals and then added back together (Eq.2.1.4):

$$\hat{z}_{\mathrm{RK}}(\mathbf{s}_0) = \mathbf{q}_0^{\mathrm{T}} \cdot \hat{\beta}_{\mathrm{GLS}} + \lambda_0^{\mathrm{T}} \cdot \mathbf{e}$$

It can be demonstrated that both KED and RK algorithms give exactly the same results (Stein, 1999; Hengl et al., 2007a). Start from KED where the predictions are made as in ordinary kriging using $\hat{z}_{\mathrm{KED}}(\mathbf{s}_0) = \lambda_{\mathrm{KED}}^{\mathrm{T}} \cdot \mathbf{z}$. The KED kriging weights ($\lambda_{\mathrm{KED}}^{\mathrm{T}}$) are obtained by solving the system (Wackernagel, 2003, p.179):

$$\begin{bmatrix} \mathbf{C} & \mathbf{q} \\ \mathbf{q}^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \lambda_{\mathrm{KED}} \\ \phi \end{bmatrix} = \begin{bmatrix} \mathbf{c_0} \\ \mathbf{q_0} \end{bmatrix}$$

where ϕ is a vector of Lagrange multipliers. Writing this out yields:

$$\begin{aligned} \mathbf{C} \cdot \lambda_{\mathrm{KED}} + \mathbf{q} \cdot \phi &= \mathbf{c_0} \\ \mathbf{q}^{\mathrm{T}} \cdot \lambda_{\mathrm{KED}} &= \mathbf{q_0} \end{aligned} \tag{2.1.27}$$

from which follows:

$$\mathbf{q}^{\mathrm{T}} \cdot \lambda_{\mathrm{KED}} = \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{c_0} - \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q} \cdot \phi \tag{2.1.28}$$

and hence:

$$\phi = \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{c_0} - \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q_0} \tag{2.1.29}$$

where the identity $\mathbf{q}^{\mathrm{T}} \cdot \lambda_{\mathrm{KED}} = \mathbf{q_0}$ has been used. Substituting ϕ back into Eq. (2.1.27) shows that the KED weights equal (Papritz and Stein, 1999, p.94):

$$\begin{aligned} \lambda_{\mathrm{KED}} &= \mathbf{C}^{-1} \cdot \mathbf{c_0} - \mathbf{C}^{-1} \cdot \mathbf{q} \cdot \left[\left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{c_0} - \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q_0}\right] \\ &= \mathbf{C}^{-1} \cdot \left[\mathbf{c_0} + \mathbf{q} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \mathbf{q} \cdot\right)^{-1} \cdot \left(\mathbf{q_0} - \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{c_0}\right)\right] \end{aligned} \tag{2.1.30}$$

Let us now turn to RK. Recall from Eq.(2.1.3) that the GLS estimate for the vector of regression coefficients is given by:

$$\hat{\beta}_{\mathrm{GLS}} = \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{z} \tag{2.1.31}$$

and weights for residuals by:

$$\lambda_0^{\mathrm{T}} = \mathbf{c}_0^{\mathrm{T}} \cdot \mathbf{C}^{-1} \tag{2.1.32}$$

and substituting these in RK formula (Eq.2.1.4) gives:

$$\begin{aligned} &= \mathbf{q}_0^{\mathrm{T}} \cdot \hat{\beta}_{\mathrm{GLS}} + \lambda_0^{\mathrm{T}} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}_{\mathrm{GLS}}) \\ &= \left[\mathbf{q}_0^{\mathrm{T}} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} + \mathbf{c}_0^{\mathrm{T}} \cdot \mathbf{C}^{-1} - \mathbf{c}_0^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1}\right] \cdot \mathbf{z} \\ &= \mathbf{C}^{-1} \cdot \left[\mathbf{c}_0^{\mathrm{T}} + \mathbf{q}_0^{\mathrm{T}} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} - \mathbf{c}_0^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}}\right] \cdot \mathbf{z} \\ &= \mathbf{C}^{-1} \cdot \left[\mathbf{c}_0 + \mathbf{q} \cdot \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \cdot \mathbf{q}\right)^{-1} \cdot \left(\mathbf{q}_0 - \mathbf{q}^{\mathrm{T}} \cdot \mathbf{C}^{-1} \mathbf{c}_0\right)\right] \cdot \mathbf{z} \end{aligned} \tag{2.1.33}$$

The left part of the equation is equal to Eq.(2.1.30), which proves that KED will give the same predictions as RK if same inputs are used. A detailed comparison of RK and KED using the 5–points example in MS Excel is also available as supplementary material[12].

Although the KED seems, at first glance, to be computationally more straightforward than RK, the variogram parameters for KED must also be estimated from regression residuals, thus requiring a separate regression modeling step. This regression should be GLS because of the likely spatial correlation between residuals. Note that many analyst use instead the OLS residuals, which may not be too different from the GLS residuals (Hengl et al., 2007a; Minasny and McBratney, 2007). However, they are not optimal if there is any spatial correlation, and indeed they may be quite different for clustered sample points or if the number of samples is relatively small ($\ll$200).

A limitation of KED is the instability of the extended matrix in the case that the covariate does not vary smoothly in space (Goovaerts, 1997, p.195). RK has the advantage that it explicitly separates trend estimation from spatial prediction of residuals, allowing the use of arbitrarily-complex forms of regression, rather than the simple linear techniques that can be used with KED (Kanevski et al., 1997). In addition, it allows the separate interpretation of the two interpolated components. For these reasons the use of the term *regression-kriging* over *universal kriging* has been advocated by the author (Hengl et al., 2007a). The emphasis on regression is important also because fitting of the deterministic part of variation is often more beneficial for the quality of final maps than fitting of the stochastic part (residuals).

2.1.5 A simple example of regression-kriging

The next section illustrates how regression-kriging computations work and compares it to ordinary kriging using the textbook example from Burrough and McDonnell (1998, p.139-141), in which five measurements are used to predict a value of the target variable (z) at an unvisited location ($\mathbf{s}_0$) (Fig. 2.6a). We extend this example by adding a hypothetical explanatory data source: a raster image of 10×10 pixels (Fig. 2.6b), which has been constructed to show a strong negative correlation with the target variable at the sample points.

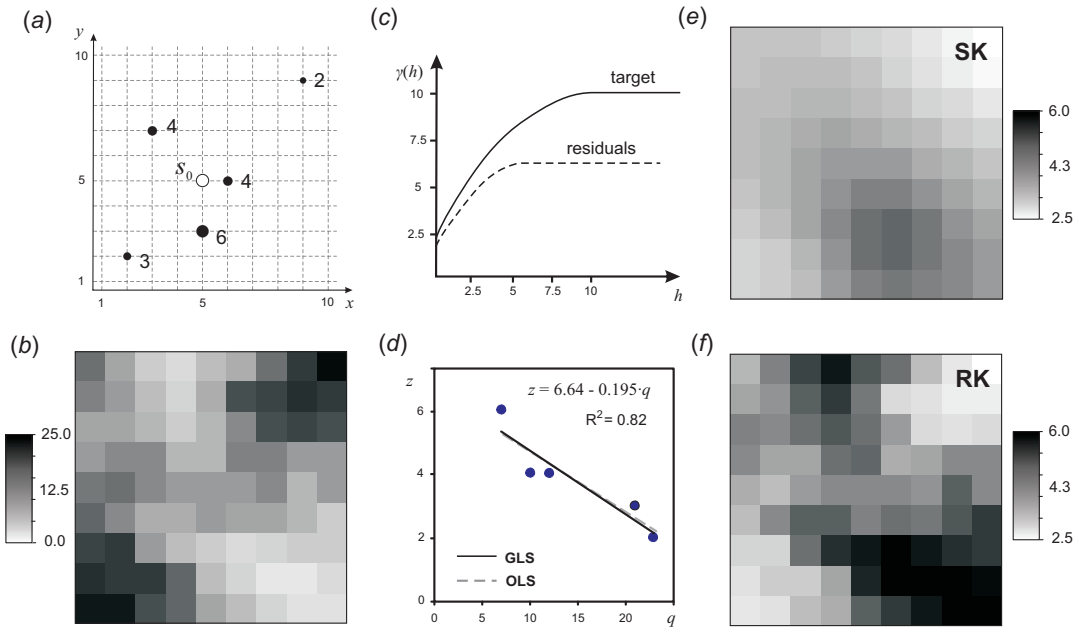

Fig. 2.6: Comparison of ordinary kriging and regression-kriging using a simple example with 5 points (Burrough and McDonnell, 1998, p.139–141): (a) location of the points and unvisited site; (b) values of the covariate q; (c) variogram for target and residuals, (d) OLS and GLS estimates of the regression model and results of prediction for a 10×10 grid using ordinary kriging (e) and regression-kriging (f). Note how the RK maps reflects the pattern of the covariate.

The RK predictions are computed as follows:

[12] http://spatial-analyst.net/book/RK5points

(1.) **Determine a linear model** of the variable as predicted by the auxiliary map q. In this case the correlation is high and negative with OLS coefficients b_0=6.64 and b_1=-0.195 (Fig. 2.6d).

(2.) **Derive the OLS residuals** at all sample locations as:

$$e^*(\mathbf{s}_i) = z(\mathbf{s}_i) - \left[b_0 + b_1 \cdot q(\mathbf{s}_i)\right] \tag{2.1.34}$$

For example, the point at (x=9, y=9) with z=2 has a prediction of $6.64 - 0.195 \cdot 23 = 1.836$, resulting in an OLS residual of $e^* = -0.164$.

(3.) **Model the covariance structure of the OLS residuals**. In this example the number of points is far too small to estimate the autocorrelation function, so we follow the original text in using a hypothetical variogram of the target variable (spherical model, nugget C_0=2.5, sill C_1=7.5 and range R=10) and residuals (spherical model, C_0=2, C_1=4.5, R=5). The residual model is derived from the target variable model of the text by assuming that the residual variogram has approximately the same form and nugget but a somewhat smaller sill and range (Fig. 2.6c), which is often found in practice (Hengl et al., 2004a).

(4.) **Estimate the GLS coefficients** using Eq.(2.1.3). In this case we get just slightly different coefficients b_0=6.68 and b_1=-0.199. The GLS coefficients will not differ much from the OLS coefficients as long there is no significant clustering of the sampling locations (Fig. 2.6d) as in this case.

(5.) **Derive the GLS residuals at all sample locations**:

$$e^{**}(\mathbf{s}_i) = z(\mathbf{s}_i) - \left[b_0 + b_1 \cdot q(\mathbf{s}_i)\right] \tag{2.1.35}$$

Note that the b now refer to the GLS coefficients.

(6.) **Model the covariance structure of the GLS residuals** as a variogram. In practice this will hardly differ from the covariance structure of the OLS residuals.

(7.) **Interpolate the GLS residuals using ordinary kriging** (OK) using the modeled variogram[13]. In this case at the unvisited point location $(5, 5)$ the interpolated residual is -0.081.

(8.) **Add the GLS surface to the interpolated GLS residuals** at each prediction point. At the unvisited point location $(5, 5)$ the explanatory variable has a value 12, so that the prediction is then:

$$\begin{aligned} \hat{z}(5,5) &= b_0 + b_1 \cdot q_i + \sum_{i=1}^{n} \lambda_i(\mathbf{s}_0) \cdot e(\mathbf{s}_i) \\ &= 6.68 - 0.199 \cdot 12 - 0.081 = 4.21 \end{aligned} \tag{2.1.36}$$

which is, in this specific case, a slightly different result than that derived by OK with the hypothetical variogram of the target variable ($\hat{z}$=4.30).

The results of OK (Fig. 2.6e) and RK (Fig. 2.6f) over the entire spatial field are quite different in this case, because of the strong relation between the covariate and the samples. In the case of RK, most of variation in the target variable (82%) has been accounted for by the predictor. Unfortunately, this version of RK has not been implemented in any software package yet[14] (see further §3.4.3). Another interesting issue is that most of the software in use (gstat, SAGA) does not estimate variogram using the GLS estimation of the residuals, but only of the OLS residuals (0 iterations). Again, for most of balanced and well spread sample sets, this will not cause any significant problems (Minasny and McBratney, 2007).

[13] Some authors argue whether one should interpolate residuals using simple kriging with zero expected mean of the residuals (by definition) or by ordinary kriging. In the case of OLS estimation, there is no difference; otherwise one should always use OK to avoid making biased estimates.

[14] Almost all geostatistical packages implement the KED algorithm because it is mathematically more elegant and hence easier to program.

2.2 Local versus localized models

In many geostatistical packages, a user can opt to limit the selection of points to determine the kriging weights by setting up a maximum distance and/or minimum and maximum number of point pairs (e.g. take only the closest 50 points). This way, the calculation of the new map can be significantly speed up. In fact, kriging in global neighborhood where $n \gg 1000$ becomes cumbersome because of computation of $\mathbf{C}^{-1}$ (Eq.1.3.5). Recall from §1.3.1 that the importance of points (in the case of ordinary kriging and assuming a standard initial variogram model) exponentially decreases with their distance from the point of interest. Typically, geostatisticians suggest that already first 30–60 closest points will be good enough to obtain stable predictions.

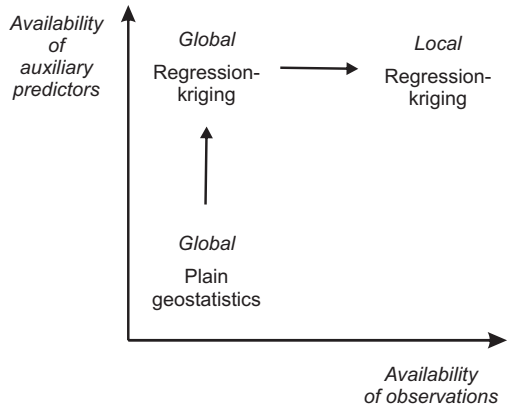

Fig. 2.7: Local regression-kriging is a further sophistication of regression-kriging. It will largely depend on the availability of explanatory and field data.

A prediction model where the search radius for derivation of kriging weights (Eq.1.3.4) is limited to a local neighborhood can be termed **localized prediction model**. There is a significant difference between *localized* and *local* prediction model, which often confuses inexperienced users. For example, if we set a search radius to re-estimate the variogram model, then we speak about a **local prediction model**, also known as the **moving window kriging** or kriging using local variograms (Haas, 1990; Walter et al., 2001; Lloyd, 2009). The local prediction model assumes that the variograms (and regression models) are non-stationary, i.e. that they need to be estimated locally.

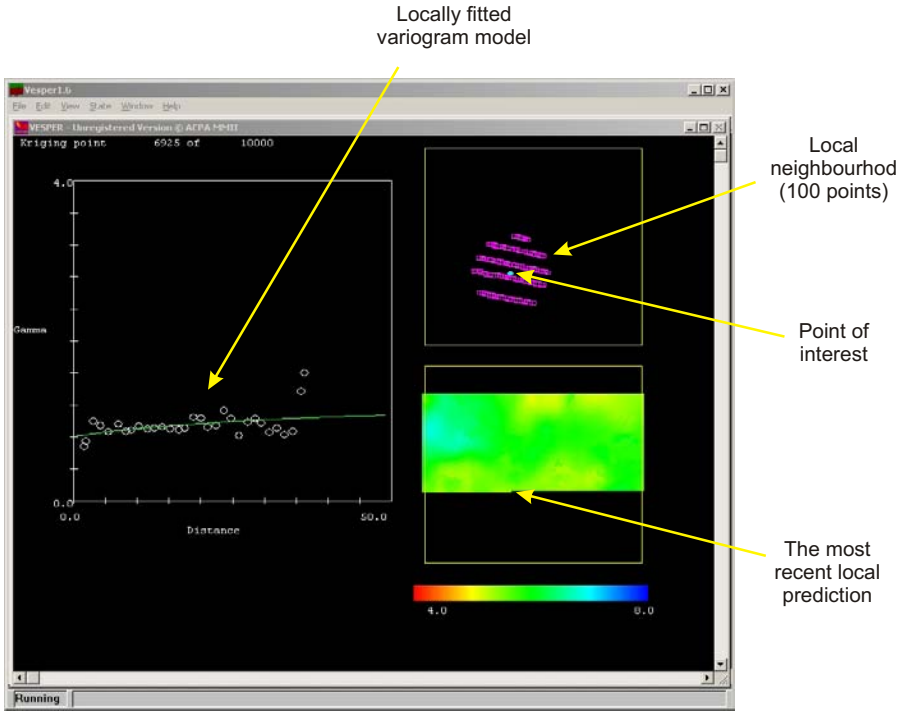

Fig. 2.8: Local variogram modeling and local ordinary kriging using a moving window algorithm in Vesper: a user can visually observe how the variograms change locally. Courtesy of Budiman Minasny.

While localized prediction models are usually just a computational trick to speed up the calculations, local prediction models are computationally much more demanding. Typically, they need to allow automated

variogram modeling and filtering of improbable models to prevent artifacts in the final outputs. A result of local prediction model (e.g. moving window variogram modeling) are not only maps of predictions, but also spatial distribution of the fitted variogram parameters (Fig. 2.7). This way we can observe how does the nugget variation changes locally, which parts of the area are smooth and which are noisy etc. Typically, local variogram modeling and prediction make sense only when we work with large point data sets (e.g. $\gg$1000 of field observations), which is still not easy to find. In addition, local variogram modeling is not implemented in many packages. In fact, the author is aware of only one: Vesper[15] (Fig. 2.8).

In the case of regression-kriging, we could also run both localized and local models. This way we will not only produce maps of variogram parameters but we would also be able to map the regression coefficients[16]. In the case of kriging with external drift, some users assume that the same variogram model can be used in various parts of the study area and limit the search window to speed up the calculations[17]. This is obviously a simplification, because in the case of KED both regression and kriging part of predictions are solved at the same time. Hence, if we limit the search window, but keep a constant variogram model, we could obtain very different predictions then if we use the global (regression-kriging) model. Only if the variogram of residuals if *absolutely* stationary, then we can limit the search window to fit the KED weights. In practice, either global (constant variogram) or local prediction models (locally estimated regression models and variograms of residuals) should be used for KED model fitting.

2.3 Spatial prediction of categorical variables

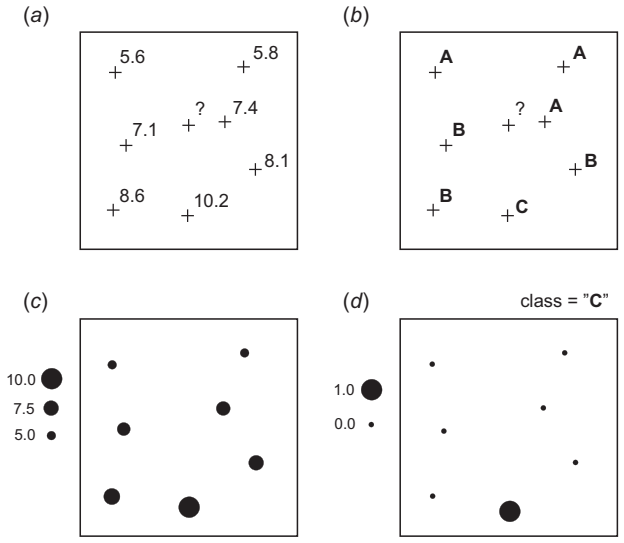

Fig. 2.9: Difficulties of predicting point-class data (b) and (d), as compared to quantitative variables (a) and (c), is that the class-interpolators are typically more complex and computationally more time-consuming.

Although geostatistics is primarily intended for use with continuous environmental variables, it can also be used to predict various types of categorical or class-type variables. Geostatistical analysis of categorical variables is by many referred to as the **indicator geostatistics** (Bierkens and Burrough, 1993). In practice, indicator kriging leads to many computational problems, which probably explains why there are not many operational applications of geostatistical mapping of categorical variables in the world (Hession et al., 2006). For example, it will typically be difficult to fit variogram for less frequent classes that occur at isolated locations (Fig. 2.9d).

Statistical grounds of indicator geostatistics has been recently reviewed by Papritz et al. (2005); Papritz (2009) who recognizes several conceptual difficulties of working with indicator data: (1) inconsistent modeling of indicator variograms, and (2) use of global variogram leads to biased predictions because the residuals are by definition non-stationary. Any attempt to use indicator kriging for data with an apparent trend either explicitly or implicitly by using ordinary indicator kriging within a local neighborhood requires the modeling of non-stationary indicator variograms to preserve the mean square optimality of kriging (Papritz, 2009). Indicator regression-kriging without any transformation has also been criticized because the model (binomial variable) suggests that residuals have mean-dependent variance ($p \cdot (1-p)$), and thus using a single variogram for the full set of residuals is not in accordance with theory.

Let us denote the field observations of a class-type variable as $z_c(\mathbf{s}_1), z_c(\mathbf{s}_2), ..., z_c(\mathbf{s}_n)$, where $c_1, c_2, ..., c_k$ are discrete categories (or states) and k is the total number of classes. A technique that estimates the soil-classes

[15] `http://www.usyd.edu.au/su/agric/acpa/vesper/vesper.html`
[16] Regression coefficients are often mapped with geographically weighted regression (Griffith, 2008).
[17] Software such as `gstat` and SAGA allow users to limit the search radius; `geoR` does not allow this flexibility.

at new unvisited location $\hat{z}_c(\mathbf{s}_0)$, given the input point data set $(z_c(\mathbf{s}_1), z_c(\mathbf{s}_2), ..., z_c(\mathbf{s}_n))$, can then be named a class-type interpolator. If spatially exhaustive predictors $q_1, q_2, ..., q_p$ (where p is the number of predictors) are available, they can be used to map each category over the area of interest. So far, there is a limited number of techniques that can achieve this:

Multi-indicator co-kriging — The simple multi-indicator kriging can also be extended to a case where several covariates are used to improve the predictions. This technique is known by the name *indicator* (soft) *co-kriging* (Journel, 1986). Although the mathematical theory is well explained (Bierkens and Burrough, 1993; Goovaerts, 1997; Pardo-Iguzquiza and Dowd, 2005), the application is cumbersome because of the need to fit a very large number of cross-covariance functions.

Multinomial Log-linear regression — This a generalization of logistic regression for situations when there are multiple classes of a target variable (Venables and Ripley, 2002). Each class gets a separate set of regression coefficients ($\beta_\mathbf{c}$). Because the observed values equal either 0 or 1, the regression coefficients need to be solved through a maximum likelihood iterative algorithm (Bailey et al., 2003), which makes the whole method somewhat more computationally demanding than simple multiple regression. An example of multinomial regression is given further in section 9.6.

Regression-kriging of indicators — One approach to interpolate soil categorical variables is to first assign memberships to point observations and then to interpolate each membership separately. This approach was first elaborated by de Gruijter et al. (1997) and then applied by Bragato (2004) and Triantafilis et al. (2001). An alternative is to first map cheap, yet descriptive, diagnostic distances and then classify these per pixel in a GIS (Carré and Girard, 2002).

In the case of logistic regression, the odds to observe a class (c) at new locations are computed as:

$$\hat{z}_c^{+}(\mathbf{s}_0) = \left[1 + \exp\left(-\beta_\mathbf{c}{}^\mathbf{T} \cdot \mathbf{q_0}\right)\right]^{-1}; \;\; c = 1, 2, .., k \tag{2.3.1}$$

where $\hat{z}_c^{+}(\mathbf{s}_0)$ are the estimated odds for class (c) at a new location s_0 and k is the number of classes. The multinomial logistic regression can also be extended to regression-kriging (for a complete derivation see Hengl et al. (2007b)). This means that the regression modeling is supplemented with the modeling of variograms for regression residuals, which can then be interpolated and added back to the regression estimate. So the predictions are obtained using:

$$\hat{z}_c^{+}(\mathbf{s}_0) = \left[1 + \exp\left(-\beta_\mathbf{c}{}^\mathbf{T} \cdot \mathbf{q_0}\right)\right]^{-1} + \hat{e}_c^{+}(\mathbf{s}_0) \tag{2.3.2}$$

where $\hat{e}_c^{+}$ are the interpolated residuals. The extension from multinomial regression to regression-kriging is not as simple as it seems. This is because the estimated values at new locations in Eq.(2.3.2) are constrained within the indicator range, which means that interpolation of residuals might lead to values outside the physical range (<0 or >1)[18]. One solution to this problem is to predict the trend part in transformed space, then interpolate residuals, sum the trend and residual part and back-transform the values (see §5.4).

Hengl et al. (2007b) show that memberships (μ_c), instead of indicators, are more suitable both for regression and geostatistical modeling, which has been also confirmed by several other authors (McBratney et al., 1992; de Gruijter et al., 1997; Triantafilis et al., 2001). Memberships can be directly linearized using the logit transformation:

$$\mu_c^{+} = \ln\left(\frac{\mu_c}{1-\mu_c}\right); \qquad 0 < \mu_c < 1 \tag{2.3.3}$$

where μ_c are the membership values used as input to interpolation. Then, all fitted values will be within the physical range (0–1). The predictions of memberships for class c at new locations are then obtained using the standard regression-kriging model (Eq.2.1.4):

$$\hat{\mu}_c^{+}(\mathbf{s}_0) = \mathbf{q_0^T} \cdot \hat{\beta}_{c,\mathtt{GLS}} + \lambda_{\mathbf{c},\mathbf{0}}^\mathbf{T} \cdot \left(\mu_\mathbf{c}^{+} - \mathbf{q} \cdot \hat{\beta}_{c,\mathtt{GLS}}\right) \tag{2.3.4}$$

[18]The degree to which they will fall outside the 0–1 range is controlled by the variogram and amount of extrapolation in feature space

The interpolated values can then be back-transformed to the membership range using (Neter et al., 1996):

$$\hat{\mu}_c(\mathbf{s}_0) = \frac{e^{\hat{\mu}_c^+(\mathbf{s}_0)}}{1 + e^{\hat{\mu}_c^+(\mathbf{s}_0)}} \tag{2.3.5}$$

In the case of regression-kriging of memberships, both spatial dependence and correlation with the predictors are modeled in a statistically sophisticated way. In addition, regression-kriging of memberships allows fitting of each class separately, which facilitates the understanding of the distribution of soil variables and the identification of problematic classes, i.e. classes which are not correlated with the predictors or do not show any spatial autocorrelation etc.

Spatial prediction of memberships can be excessive in computation time. Another problem is that, if the interpolated classes (odds, memberships) are fitted only by using the sampled data, the predictions of the odds/memberships will commonly not sum to unity at new locations. In this case, one needs to standardize values for each grid node by diving the original values by the sum of odds/memberships to ensure that they sum to unity, which is an *ad-hoc* solution. An algorithm, such as compositional regression-kriging[19] will need to be developed.

A number of alternative hybrid class-interpolators exists, e.g. the Bayesian Maximum Entropy (BME) approach by D'Or and Bogaert (2005). Another option is to use Markov-chain algorithms (Li et al., 2004, 2005a). However, note that although use of the BME and Markov-chain type of algorithms is a promising development, their computational complexity makes it still far from use in operational mapping.

2.4 Geostatistical simulations

Regression-kriging can also be used to generate simulations of a target variable using the same inputs as in the case of spatial prediction system. An equiprobable realization of an environmental variable can be generated by using the sampled values and their variogram model:

$$Z^{(\mathtt{SIM})}(\mathbf{s}_0) = E\left\{Z|z(\mathbf{s}_j), \gamma(\mathbf{h})\right\} \tag{2.4.1}$$

where $Z^{(\mathtt{SIM})}$ is the simulated value at the new location. The most common technique in geostatistics that can be used to generate equiprobable realizations is the **Sequential Gaussian Simulation** (Goovaerts, 1997, p.380-392). It starts by defining a random path for visiting each node of the grid once. At first node, kriging is used to determine the location-specific mean and variance of the conditional cumulative distribution function. A simulated value can then be drawn by using the inverse normal distribution (Box and Muller, 1958; Banks, 1998):

$$z_i^{\mathtt{SIM}} = \hat{z}_i + \hat{\sigma}_i \cdot \sqrt{-2 \cdot \ln(1-A)} \cdot \cos(2 \cdot \pi \cdot B) \tag{2.4.2}$$

where $z_i^{\mathtt{SIM}}$ is the simulated value of the target variable with induced error, A and B are the independent random numbers within the $0-0.99\ldots$ range, $\hat{z}_i$ is the estimated value at ith location, and $\hat{\sigma}_i$ is the regression-kriging error. The simulated value is then added to the original data set and the procedure is repeated until all nodes have been visited. Geostatistical simulations are used in many different fields to generate multiple realizations of the same feature (Heuvelink, 1998; Kyriakidis et al., 1999), or to generate realistic visualizations of a natural phenomena (Hengl and Toomanian, 2006; Pebesma et al., 2007). Examples of how to generate geostatistical simulations and use them to estimate the propagated error are further shown in section 10.3.2.

2.5 Spatio-temporal regression-kriging

In statistics, temporal processes (time series analysis, longitudinal data analysis) are well-known, but mixed spatio-temporal processes are still rather experimental (Banerjee et al., 2004). The 2D space models can be

[19] Walvoort and de Gruijter (2001), for example, already developed a compositional solution for ordinary kriging that will enforce estimated values to sum to unity at all locations.

extended to the time domain, which leads to **spatio-temporal geostatistics** (Kyriakidis and Journel, 1999). The universal kriging model (Eq.2.1.1) then modifies to:

$$Z(\mathbf{s},t) = m(\mathbf{s},t) + \varepsilon'(\mathbf{s},t) + \varepsilon'' \tag{2.5.1}$$

where $\varepsilon'(\mathbf{s},t)$ is the spatio-temporally autocorrelated residual for every $(\mathbf{s},t) \in S \times T$, while $m(\mathbf{s},t)$, the deterministic component of the model, can be estimated using e.g. (Fassó and Cameletti, 2009):

$$m(\mathbf{s},t) = \mathbf{q}(\mathbf{s},t) \cdot \beta + \mathbf{K}(s) \cdot \mathbf{y}_t + \omega(\mathbf{s},t) \tag{2.5.2}$$

where $\mathbf{q}$ is a matrix of covariates available at all $\mathbf{s},t$ locations, $\mathbf{y_t}$ is a component of a target variable that is constant in space (global trend), $\mathbf{K}(s)$ is a matrix of coefficients, and $\omega(\mathbf{s},t)$ is the spatial small-scale component (white noise in time) correlated over space.

A possible but tricky simplification of the space-time models is to consider time to be third dimension of space. In that case, spatio-temporal interpolation follows the same interpolation principle as explained in Eq.(1.1.2), except that here the variograms are estimated in three dimensions (two-dimensional position x and y and '*position*' in time). From the mathematical aspect, the extension from the static 2D interpolation to the 3D interpolation is then rather simple. Regression modeling can be simply extended to a space-time model by adding time as a predictor. For example, a spatio-temporal regression model for interpolation of land surface temperature (see further §2.9.2) would look like this:

$$\begin{aligned} LST(\mathbf{s}_0,t_0) = {} & b_0 + b_1 \cdot DEM(\mathbf{s}_0) + b_2 \cdot LAT(\mathbf{s}_0) + b_3 \cdot DISTC(\mathbf{s}_0) + b_4 \cdot LSR(\mathbf{s}_0,t_0) \\ & + b_5 \cdot SOLAR(\mathbf{s}_0,t_0) + b_6 \cdot \cos\left([t_0 - \phi] \cdot \frac{\pi}{180}\right); \qquad \Delta t = 1 \text{ day} \end{aligned} \tag{2.5.3}$$

where *DEM* is the elevation map, *LAT* is the map showing distance from the equator, *DISTC* is the distance from the coast line, *LSR* is the land surface radiation from natural or man-made objects, *SOLAR* is the direct solar insolation for a given cumulative Julian day $t \in (0,+\infty)$, $\cos(t)$ is a generic function to account for seasonal variation of values and ϕ is the phase angle[20]. *DEM*, *LAT*, *DISTC* are temporally-constant predictors, while surface radiation and solar insolation maps need to be provided for each time interval used for data fitting.

The residuals from this regression model can then be analyzed for (spatio-temporal) auto-correlation. In `gstat`, extension from 2D to 3D variograms is possible by extending the variogram parameters: for 3D space-time variograms five values should be given in the form `anis = c(p,q,r,s,t)`, where `p` is the angle for the **principal direction of continuity** (measured in degrees, clockwise from y, in direction of x), `q` is the **dip angle** for the principal direction of continuity (measured in positive degrees up from horizontal), `r` is the third rotation angle to rotate the two minor directions

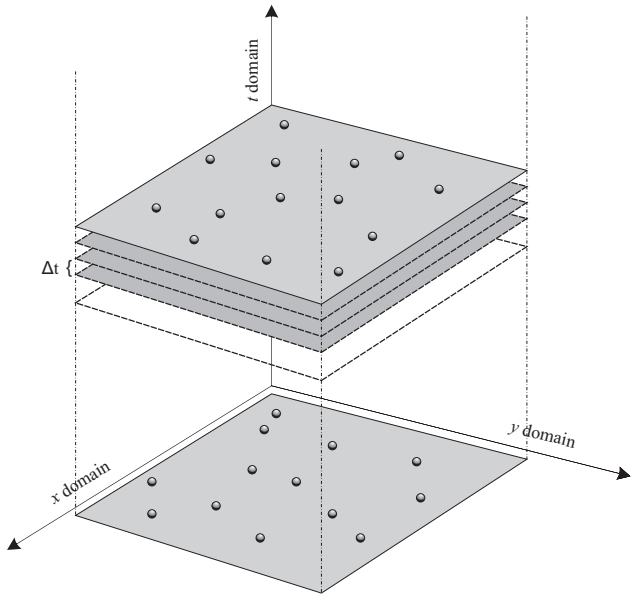

Fig. 2.10: Extension of a 2D prediction model to the space-time domain. Note that in the space-time cube, the amount of pixels needed to store the data exponentially increases as a function of: width × height × number of predictors × number of time intervals.

[20] A time delay from the coldest day.

around the principal direction defined by p and q[21] (see Fig. 1.11). A positive angle acts counter-clockwise while looking in the principal direction.

Once we have fitted the space-time variogram, we can run regression-kriging to estimate the values at 3D locations. In practice, we only wish to produce maps for a given time interval (t_0=constant), i.e. to produce 2D-slices of values in time (Fig. 2.10). Once we have produced a time-series of predictions, we can analyze the successive time periods and run various types of time-series analysis. This will help us detect temporal trends spatially and extract informative images about the dynamics of the feature of interest.

Note that, in order to yield accurate predictions using spatio-temporal techniques, dense sampling in both space and time is required. This means that existing natural resource surveys that have little to no repetition in time ($\ll$10 repetitions in time) cannot be adopted. Not to mention the computational complexity as the maps of predictors now multiply by the amount of time intervals. In addition, estimation of the spatio-temporal variograms will often be a cumbersome because we need to fit space-time models, for which we might not have enough space-time observations. A review of spatio-temporal models, i.e. dynamic linear state-space models, and some practical suggestions how to analyze such data and fit spatially varying coefficients can be followed in Banerjee et al. (2004, §8).

A specific extension of the general model from Eq.(2.5.1) is to estimate the deterministic part of variation by using process-based (simulation) models, which are often based on differential equations. In this case an environmental variable is predicted from a set of environmental predictors incorporated in a dynamic model (Eq.1.3.12):

$$Z(\mathbf{s},t) = f_{s,c,r,p,a}(t) + \varepsilon'(\mathbf{s},t) + \varepsilon'' \tag{2.5.4}$$

where s, c, r, p, a are the input (zero-stage) environmental conditions and f is a mathematical deterministic function that can be used to predict the values for a given space-time position. This can be connected with the Einstein's assumption that the Universe is in fact a trivial system that can be modeled and analyzed using "a one–dimensional differential equation — in which everything is a function of time"[22]. Some examples of operational soil-landscape process-based models are given by Minasny and McBratney (2001) and Schoorl et al. (2002). In vegetation science, for example, global modeling has proven to be very efficient for explanation of the actual distribution of vegetation and of global changes (Bonan et al., 2003). Integration of environmental process-based models will soon lead to development of a global dynamic model of environmental systems that would then provide solutions for different multipurpose national or continental systems.

Fassó and Cameletti (2009) recently proposed hierarchical models as a general approach for spatio-temporal problems, including dynamical mapping, and the analysis of the outputs from complex environmental modeling chains. The hierarchical models are a suitable solution to spatio-temporal modeling because they make it possible to define the joint dynamics and the full likelihood; the maximum likelihood estimation can be further simplified by using Expectation-Maximization algorithm. The basis of this approach is the classical two-stage hierarchical state-space model (Fassó and Cameletti, 2009):

$$Z_t = \mathbf{q}_t \cdot \beta + \mathbf{K} \cdot \mathbf{y}_t + \mathbf{e}_t \tag{2.5.5}$$

$$\mathbf{y}_t = \mathbf{G} \cdot \mathbf{y}_{t-1} + \eta_t \tag{2.5.6}$$

where $\mathbf{y}_t$ is modeled as the autoregressive process, $\mathbf{G}$ is the transition matrix and η_t is the innovation error. If all parameters are known, the unobserved temporal process $\mathbf{y}_t$ can be estimated for each time point t using e.g. *Kalman filter* or *Kalman smoother*. Such process-based spatio-temporal models are still experimental and it make take time until their semi-automated software implementations appear in R.

2.6 Species Distribution Modeling using regression-kriging

The key inputs to a Species Distribution Model (SDM) are: the inventory (population) of animals or plants consisting of a total of N individuals (a point pattern $\mathbf{X} = \{\mathbf{s}_i\}_1^N$; where $\mathbf{s}_i$ is a spatial location of individual

[21] http://www.gstat.org/manual/node20.html

[22] Quote by James Peebles, Princeton, 1990; published in "God's Equation: Einstein, Relativity, and the Expanding Universe" by Amir D. Aczel.

animal or plant; Fig. 1.3a), covering some area $B_{HR} \subset \mathbb{R}^2$ (where *HR* stands for home-range and $\mathbb{R}^2$ is the Euclidean space), and a list of environmental covariates/predictors $(q_1, q_2, \ldots q_p)$ that can be used to explain spatial distribution of a target species. In principle, there are two distinct groups of statistical techniques that can be used to map the realized species' distribution: (a) the point pattern analysis techniques, such as kernel smoothing, which aim at predicting density of a point process (Fig. 2.11a); and (b) statistical, GLM-based, techniques that aim at predicting the probability distribution of occurrences (Fig. 2.11c). Both approaches are explained in detail in the following sections.

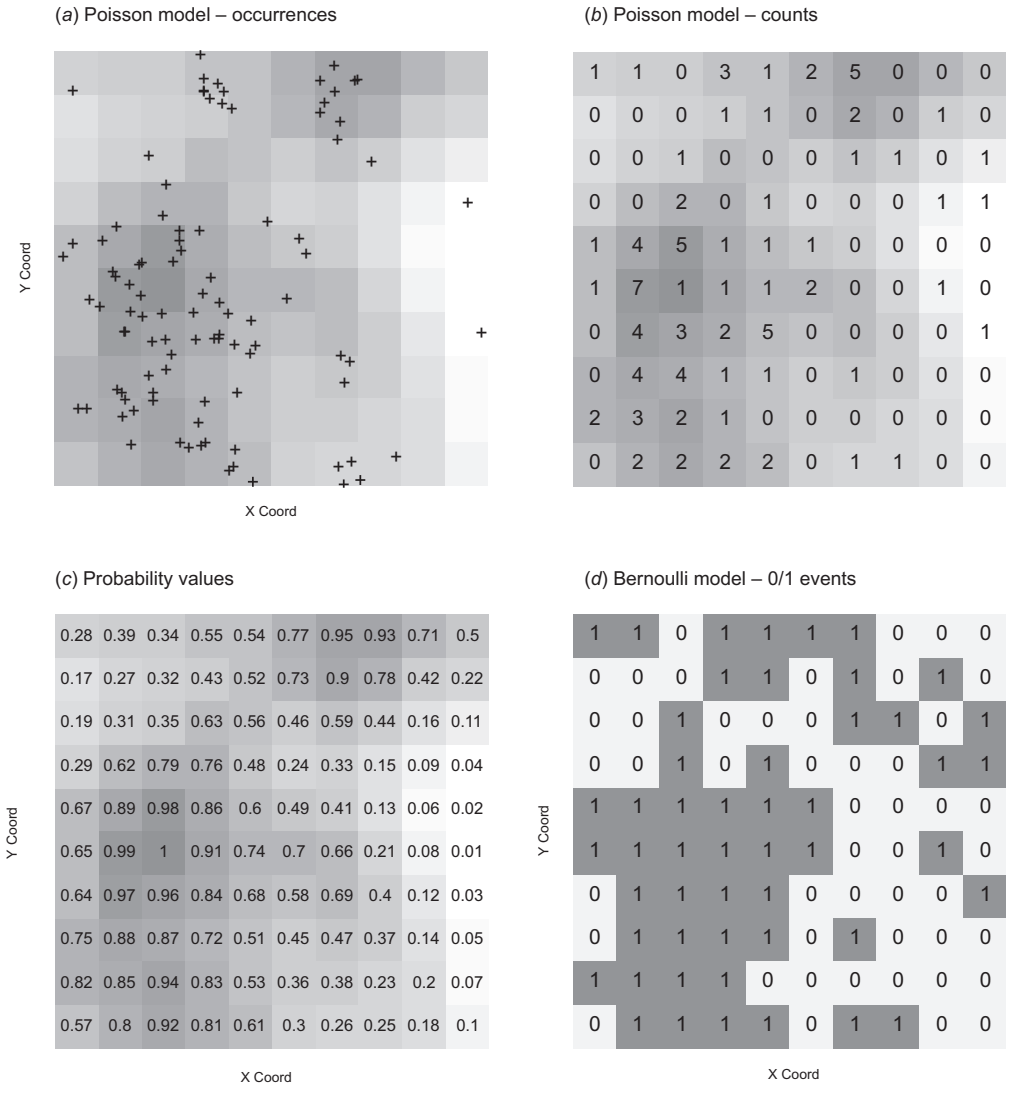

Fig. 2.11: Examples of (simulated) species distribution maps produced using common statistical models.

Species' density estimation using kernel smoothing and covariates

Spatial density (ν; if unscaled, also known as *"spatial intensity"*) of a point pattern (ignoring the time dimension) is estimated as:

$$\mathbb{E}\left[N(\mathbf{X} \cap B)\right] = \int_B \nu(\mathbf{s}) d\mathbf{s} \tag{2.6.1}$$

In practice, it can be estimated using e.g. a kernel estimator (Diggle, 2003; Baddeley, 2008):

$$\nu(\mathbf{s}) = \sum_{i=1}^{n} \kappa \cdot \left(\left\| \mathbf{s} - \mathbf{s}_i \right\| \right) \cdot b(\mathbf{s}) \tag{2.6.2}$$

where $\nu(\mathbf{s})$ is spatial density at location $\mathbf{s}$, $\kappa(\mathbf{s})$ is the kernel (an arbitrary probability density), $\mathbf{s}_i$ is location of an occurrence record, $\left\| \mathbf{s} - \mathbf{s}_i \right\|$ is the distance (norm) between an arbitrary location and observation location, and $b(\mathbf{s})$ is a border correction to account for missing observations that occur when $\mathbf{s}$ is close to the border of the region (Fig. 2.11a). A common (isotropic) kernel estimator is based on a Gaussian function with mean zero and variance 1:

$$\widehat{\nu}(\mathbf{s}) = \frac{1}{H^2} \cdot \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\|\mathbf{s}-\mathbf{s}_i\|^2}{2}} \cdot b(\mathbf{s}) \tag{2.6.3}$$

The key parameter for kernel smoothing is the bandwidth (H) i.e. the smoothing parameter, which can be connected with the choice of variogram in geostatistics. The output of kernel smoothing is typically a map (raster image) consisting of M grid nodes, and showing spatial pattern of species' clustering.

Spatial density of a point pattern can also be modeled using a list of spatial covariates q's (in ecology, we call this environmental predictors), which need to be available over the whole area of interest B. For example, using a Poisson model (Baddeley, 2008):

$$\log \nu(\mathbf{s}) = \log \beta_0 + \log q_1(\mathbf{s}) + \ldots + \log q_p(\mathbf{s}) \tag{2.6.4}$$

where log transformation is used to account for the skewed distribution of both density values and covariates; p is the number of covariates. Models with covariates can be fitted to point patterns e.g. in the `spatstat` package[23]. Such point pattern–covariates analysis is commonly run only to determine i.e. test if the covariates are correlated with the feature of interest, to visualize the predicted trend function, and to inspect the spatial trends in residuals. Although statistically robust, point pattern–covariates models are typically not considered as a technique to improve prediction of species' distribution. Likewise, the model residuals are typically not used for interpolation purposes.

Predicting species' distribution using ENFA and GLM (pseudo-absences)

An alternative approach to spatial prediction of species' distribution using occurrence-only records and environmental covariates is the combination of ENFA and regression modeling. In general terms, predictions are based on fitting a GLM:

$$\mathbb{E}(\mathbf{P}) = \mu = g^{-1}(\mathbf{q} \cdot \beta) \tag{2.6.5}$$

where E($\mathbf{P}$) is the expected probability of species occurrence ($P \in [0,1]$; Fig. 2.11c), $\mathbf{q} \cdot \beta$ is the linear regression model, and g is the link function. A common link function used for SDM with presence observations is the logit link function:

$$g(\mu) = \mu^{+} = \ln\left(\frac{\mu}{1-\mu}\right) \tag{2.6.6}$$

and the Eq.(2.6.5) becomes logistic regression (Kutner et al., 2004).

The problem of running regression analysis with occurrence-only observations is that we work with 1's only, which means that we cannot fit any model to such data. To account for this problem, species distribution modelers (see e.g. Engler et al. (2004); Jiménez-Valverde et al. (2008) and Chefaoui and Lobo (2008)) typically insert the so-called "*pseudo-absences*" — 0's simulated using a plausible models, such as **Environmental**

[23] This actually fits the maximum pseudolikelihood to a point process; for more details see Baddeley (2008).

Niche Factor Analysis (ENFA), MaxEnt or GARP (Guisan and Zimmermann, 2000), to depict areas where a species is not likely to occur. ENFA is a type of factor analysis that uses observed presences of a species to estimate which are the most favorable areas in the feature space, and then uses this information to predict the potential distribution of species for all locations (Hirzel and Guisan, 2002). The difference between ENFA and the Principal Component Analysis is that the ENFA factors have an ecological meaning. ENFA results in a Habitat Suitability Index (HSI$\in [0-100\%]$) — by depicting the areas of low HSI, we can estimate where the species is very unlikely to occur, and then simulate a new point pattern that can be added to the occurrence locations to produce a '*complete*' occurrences+absences data set. Once we have both 0's and 1's, we can fit a GLM as shown in Eq.(2.6.5) and generate predictions (probability of occurrence) using geostatistical techniques as described in e.g. Gotway and Stroup (1997).

Predicting species' density using ENFA and logistic regression-kriging

Point pattern analysis, ENFA and regression-kriging can be successfully combined using the approach explained in Hengl et al. (2009b). First, we will assume that our input point pattern represents only a sample of the whole population ($\mathbf{X}_S = \{\mathbf{s}_i\}_1^n$), so that the density estimation needs to be standardized to avoid biased estimates. Second, we will assume that pseudo-absences can be generated using both information about the potential habitat (HSI) and geographical location of the occurrence-only records. Finally, we focus on mapping the actual count of individuals over the grid nodes (realized distribution), instead of mapping the probability of species' occurrence.

Spatial density values estimated by kernel smoothing are primarily controlled by the bandwidth size (Bivand et al., 2008). The higher the bandwidth, the lower the values in the whole map; likewise, the higher the sampling intensity (n/N), the higher the spatial density, which eventually makes it difficult to physically interpret mapped values. To account for this problem, we propose to use relative density ($\nu_r : B \rightarrow [0,1]$) expressed as the ratio between the local and maximum density at all locations:

$$\nu_r(\mathbf{s}) = \frac{\nu(\mathbf{s})}{\max\{\nu(\mathbf{s})|\mathbf{s} \in B\}_1^M} \tag{2.6.7}$$

An advantage of using the relative density is that the values are in the range $[0,1]$, regardless of the bandwidth and sample size (n/N). Assuming that our sample $\mathbf{X}_S$ is representative and unbiased, it can be shown that $\nu_r(\mathbf{s})$ is an unbiased estimator of the true spatial density (see e.g. Diggle (2003) or Baddeley (2008)). In other words, regardless of the sample size, by using relative intensity we will always be able to produce an unbiased estimator of the spatial pattern of density for the whole population (see further Fig. 8.4).

Furthermore, assuming that we actually know the size of the whole population (N), by using predicted relative density, we can also estimate the actual spatial density (number of individuals per grid node; as shown in Fig. 2.11b):

$$\nu(\mathbf{s}) = \nu_r(\mathbf{s}) \cdot \frac{N}{\sum_{j=1}^M \nu_r(\mathbf{s})}; \qquad \sum_{j=1}^M \nu(\mathbf{s}) = N \tag{2.6.8}$$

which can be very useful if we wish to aggregate the species' distribution maps over some polygons of interest, e.g. to estimate the actual counts of individuals.

Our second concern is the insertion of pseudo-absences. Here, two questions arise: (1) how many pseudo-absences should we insert? and (b) where should we locate them? Intuitively, it makes sense to generate the same number of pseudo-absence locations as occurrences. This is also supported by the statistical theory of model-based designs, also known as "*D-designs*". For example, assuming a linear relationship between density and some predictor q, the optimal design that will minimize the prediction variance is to put half of observation at one extreme and other at other extreme. All D-designs are in fact symmetrical, and all advocate higher spreading in feature space (for more details about D-designs, see e.g. Montgomery (2005)), so this principle seems logical. After the insertion of the pseudo-absences, the extended observations data set is:

$$\mathbf{X_f} = \left\{\{\mathbf{s}_i\}_1^n, \{\mathbf{s^*}_i\}_1^{n*}\right\}; \qquad n = n^* \tag{2.6.9}$$

where $\mathbf{s}^*_i$ are locations of the simulated pseudo-absences. This is not a point pattern any more because now also quantitative values — either relative densities ($\nu_r(\mathbf{s}_i)$) or indicator values — are attached to locations ($\mu(\mathbf{s}_i) = 1$ and $\mu(\mathbf{s}^*_i) = 0$).

The remaining issue is where and how to allocate the pseudo-absences? Assuming that a spreading of species in an area of interest is a function of the potential habitat and assuming that the occurrence locations on the HSI axis will commonly be skewed toward high values (see further Fig. 8.8 left; see also Chefaoui and Lobo (2008)), we can define the probability distribution (τ) to generate the pseudo-absence locations as e.g.:

$$\tau(\mathbf{s}^*) = [100\% - \text{HSI}(\mathbf{s})]^2 \tag{2.6.10}$$

where the square term is used to insure that there are progressively more pseudo-absences at the edge of low HSI. This way also the pseudo-absences will approximately follow Poisson distribution. In this paper we propose to extend this idea by considering location of occurrence points in geographical space also (see also an interesting discussion on the importance of geographic extent for generation of pseudo-absences by VanDerWal et al. (2009)). The Eq.(2.6.10) then modifies to:

$$\tau(\mathbf{s}^*) = \left[\frac{d_R(\mathbf{s}) + (100\% - \text{HSI}(\mathbf{s}))}{2}\right]^2 \tag{2.6.11}$$

where d_R is the normalized distance in the range $[0, 100\%]$, i.e. the distance from the observation points ($\mathbf{X}$) divided by the maximum distance. By using Eq.(2.6.11) to simulate the pseudo-absence locations, we will purposively locate them both geographically further away from the occurrence locations and in the areas of low HSI (unsuitable habitat).

After the insertion of pseudo-absences, we can attach to both occurrence-absence locations values of estimated relative density, and then correlate this with environmental predictors. This now becomes a standard geostatistical point data set, representative of the area of interest, and with quantitative values attached to point locations (see further Fig. 8.10d). Recall from Eq.(2.6.7) that we attach relative intensities to observation locations. Because these are bounded in the $[0, 1]$ range, we can use the logistic regression model to make predictions. Thus, the relative density at some new location ($\mathbf{s}_0$) can be estimated using:

$$\widehat{\nu}^+_r(\mathbf{s}_0) = \left[1 + \exp\left(-\beta^{\mathbf{T}} \cdot \mathbf{q_0}\right)\right]^{-1} \tag{2.6.12}$$

where β is a vector of fitted regression coefficients, $\mathbf{q_0}$ is a vector of predictors (maps) at a new location, and $\widehat{\nu}^+_r(\mathbf{s}_0)$ is the predicted logit-transformed value of the relative density. Assuming that the sampled intensities are continuous values in the range $\nu_r \in (0, 1)$, the model in Eq.(2.6.12) is in fact a liner model, which allows us to extend it to a more general linear geostatistical model such as regression-kriging. This means that the regression modeling is supplemented with the modeling of variograms for regression residuals, which can then be interpolated and added back to the regression estimate (Eq.2.1.4):

$$\widehat{\nu}^+_r(\mathbf{s}_0) = \mathbf{q_0^T} \cdot \hat{\beta}_{\text{GLS}} + \delta_{\mathbf{0}}^{\mathbf{T}} \cdot \left(\nu^+_{\mathbf{r}} - \mathbf{q} \cdot \hat{\beta}_{\text{GLS}}\right) \tag{2.6.13}$$

where $\delta_{\mathbf{0}}$ is the vector of fitted weights to interpolate the residuals using ordinary kriging. In simple terms, logistic regression-kriging consists of five steps:

(1.) convert the relative intensities to logits using Eq.(2.6.6); if the input values are equal to 0/1, replace with the second smallest/highest value;

(2.) fit a linear regression model using Eq.(2.6.12);

(3.) fit a variogram for the residuals (logits);

(4.) produce predictions by first predicting the regression-part, then interpolate the residuals using ordinary kriging; finally add the two predicted trend-part and residuals together (Eq.2.6.13)

(5.) back-transform interpolated logits to the original $(0, 1)$ scale by:

$$\widehat{v}_r(\mathbf{s}_0) = \frac{e^{\widehat{v}_r^{+}(\mathbf{s}_0)}}{1 + e^{\widehat{v}_r^{+}(\mathbf{s}_0)}} \tag{2.6.14}$$

After we have mapped relative density over area of interest, we can also estimate the actual counts using the Eq.(2.6.8). This procedure is further elaborated in detail in chapter 8.

2.7 Modeling of topography using regression-kriging

A **Digital Elevation Model** (DEM) is a digital representation of the land surface — the major input to quantitative analysis of topography, also known as Digital Terrain Analysis or Geomorphometry (Wilson and Gallant, 2000; Hengl and Reuter, 2008). Typically, a DEM is a raster map (an image or an elevation array) that, like many other spatial features, can be efficiently modeled using geostatistics. The geostatistical concepts were introduced in geomorphometry by Fisher (1998) and Wood and Fisher (1993), then further elaborated by Kyriakidis et al. (1999), Holmes et al. (2000) and Oksanen (2006). An important focus of using geostatistics to model topography is assessment of the errors in DEMs and analysis of effects that the DEM errors have on the results of spatial modeling. This is the principle of error propagation that commonly works as follows: simulations are generated from point-measured heights to produce multiple equiprobable realizations of a DEM of an area; a spatial model is applied m times and output maps then analyzed for mean values and standard deviations per pixel; the results of analysis can be used to quantify DEM accuracy and observe impacts of uncertain information in various parts of the study area (Hunter and Goodchild, 1997; Heuvelink, 1998; Temme et al., 2008).

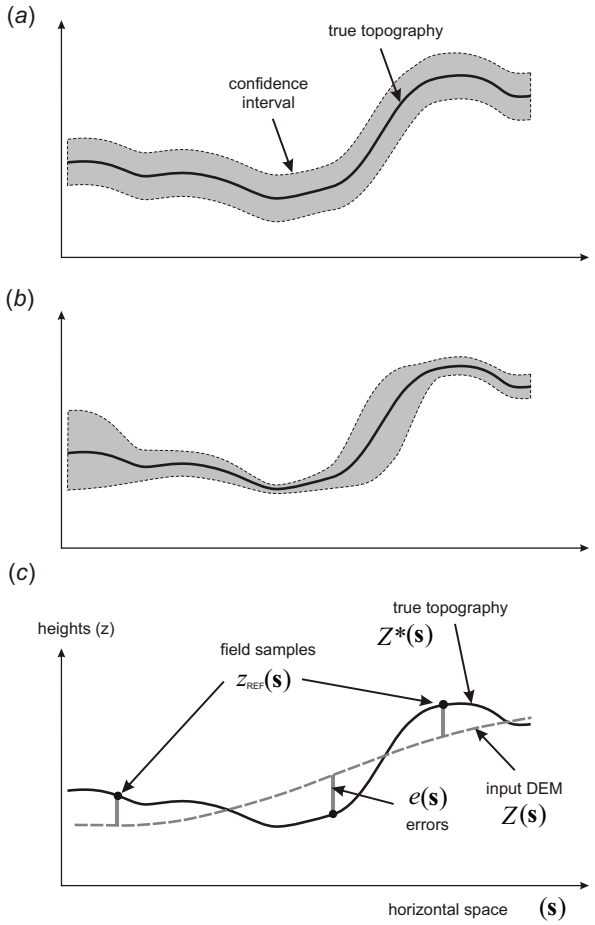

Fig. 2.12: Conceptual aspects of modeling topography using geostatistics. A cross section showing the true topography and the associated uncertainty: (a) constant, global uncertainty model and (b) spatially variable uncertainty; (c) estimation of the DEM errors using precise height measurements.

So far, DEMs have been modeled by using solely point-sampled elevations. For example, ordinary kriging is used to generate DEMs (Mitas and Mitasova, 1999; Lloyd and Atkinson, 2002); conditional geostatistical simulations are used to generate equiprobable realizations of DEMs (Fisher, 1998; Kyriakidis et al., 1999). In most studies, no explanatory information on topography is employed directly in the geostatistical modeling. Compared to the approach of Hutchinson (1989, 1996) where auxiliary maps of streams are often used to produce hydrologically-correct DEMs, the geostatistical approach to modeling of topography has often been limited to analysis of point-sampled elevations.

2.7.1 Some theoretical considerations

DEMs are today increasingly produced using automated (mobile GPS) field sampling of elevations or airborne scanning devices (radar or LiDAR-based systems). In the case elevations are sampled at sparsely-located points,

a DEM can be generated using geostatistical techniques such as ordinary kriging (Wood and Fisher, 1993; Mitas and Mitasova, 1999). The elevation at some grid node ($\mathbf{s}_0$) of the output DEM can be interpolated using ordinary kriging (Eq.1.3.2); the same technique can be used to produce simulated DEMs (see section 2.4). Direct simulation of DEMs using the sampled elevations is discussed in detail by Kyriakidis et al. (1999).

The use of kriging in geomorphometry to generate DEMs has been criticized by many (Wood and Fisher, 1993; Mitas and Mitasova, 1999; Li et al., 2005b), mainly because it leads to many artifacts, it oversmooths elevations and it is very sensitive to sampling density and local extreme values. So far, splines have been used in geomorphometry as a preferred technique to generate DEMs or to filter local errors (Mitasova et al., 2005). More recently, Hengl et al. (2008) demonstrated that regression-kriging can be used to employ auxiliary maps, such as maps of drainage patterns, land cover and remote sensing-based indices, directly in the geostatistical modeling of topography. Details are now discussed in the succeeding sections.

If additional, auxiliary maps (drainage network, water bodies, physiographic break-lines) are available, a DEM can be generated from the point-measured elevations using the regression-kriging model (Eq.2.1.4). The biggest advantage of using auxiliary maps is a possibility to more precisely model uncertainty of the sampled elevations and analyze which external factors cause this variability. Whereas, in pure statistical Monte Carlo approach where we work with global, constant parameters (Fig. 2.12a), in the case of geostatistical modeling, the DEM uncertainty can be modeled with a much higher level of detail (Fig. 2.12b).

In the case a DEM is obtained from an airborne or satellite-based scanning mission (radar, LiDAR or stereoscopic images), elevations are already available over the whole area of interest. Geostatistics is then used to analyze inherent errors in the DEM images (Grohmann, 2004), filter local errors caused by physical limitations of the instrument (Lloyd and Atkinson, 2002; Evans and Hudak, 2007), and eventually cluster the area according to their statistical properties (Lloyd and Atkinson, 1998).

Geostatistical simulation of complete elevation data is somewhat more complicated than with point data. At the moment, the simulations of DEM images are most commonly obtained by simulating error surfaces derived from additional field-control samples (Fig. 2.12c). The elevations measured at control points are used to assess the errors. The point map of DEM errors can then be used to generate equiprobable error surfaces, which are then added to the original DEM to produce an equiprobable realization of a DEM (Hunter and Goodchild, 1997; Holmes et al., 2000; Endreny and Wood, 2001; Temme et al., 2008). From a statistical perspective, a DEM produced directly by using scanning devices (SRTM, LiDAR) consists of three components: $Z^*(\mathbf{s})$ the deterministic component, $\varepsilon'(\mathbf{s})$ the spatially correlated random component, and ε'' is the pure noise, usually the result of the measurement error. In raster-GIS terms, we can decompose a DEM into two grids: (1) the deterministic DEM and (2) the error surface. If precise point-samples of topography (e.g. highly precise GPS measurements) are available, they can be used to estimate the errors (Fig. 2.12c):

$$e(\mathbf{s}_i) = z^*_{\mathsf{REF}}(\mathbf{s}_i) - Z(\mathbf{s}_i); \quad E\left\{e(\mathbf{s})\right\} = 0 \tag{2.7.1}$$

The measured errors at point locations can also be manipulated using geostatistics to generate the **error surface**:

$$e^{(\mathsf{SIM})}(\mathbf{s}_0) = E\left\{\varepsilon | e(\mathbf{s}_i), \gamma_e(\mathbf{h})\right\} \tag{2.7.2}$$

The simulated error surface can then be added to the deterministic DEM to produce an equiprobable realization of a DEM:

$$z^{(\mathsf{SIM})}(\mathbf{s}_j) = z^*(\mathbf{s}_j) + e^{(\mathsf{SIM})}(\mathbf{s}_j) \tag{2.7.3}$$

An obvious problem with this approach is that the deterministic DEM ($z^*(\mathbf{s}_j)$) is usually not available, so that the input DEM is in fact used to generate simulations, which leads to (see e.g. Holmes et al. (2000); Temme et al. (2008)):

$$\begin{aligned} z^{(\mathsf{SIM})}(\mathbf{s}_j) &= z(\mathbf{s}_j) + e^{(\mathsf{SIM})}(\mathbf{s}_j) \\ &= z^*(\mathbf{s}_j) + \varepsilon'(\mathbf{s}_j) + \varepsilon'' + e^{(\mathsf{SIM})}(\mathbf{s}_j) \end{aligned} \tag{2.7.4}$$

which means that the simulated error surface and the inherent error component, at some locations, will double, and at others will annul each other. However, because the inherent error and the simulated error are in fact independent, the mean of the summed errors will be close to zero (unbiased simulation), but the standard deviation of the error component will be on average 40% larger. Hence a DEM simulated using Eq.(2.7.3) will be much noisier than the original DEM. The solution to this problem is to substitute the deterministic DEM component with a smoother DEM, e.g. a DEM derived from contour lines digitized from a finer-scale topo-map. As an alternative, the deterministic DEM component can be prepared by smoothing the original DEM i.e. filtering it for known noise and systematic errors (see e.g. Selige et al. (2006)).

2.7.2 Choice of auxiliary maps

The spatially correlated error component will also often correlate with the explanatory information (Oksanen, 2006). For example, in traditional cartography, it is known that the error of measuring elevations is primarily determined by the complexity of terrain (the slope factor), land cover (density of objects) and relative visibility (the shadow effect). Especially in the cases where the DEMs are produced through photogrammetric methods, information about the terrain shading can be used to estimate the expected error of measuring heights. Similarly, a SRTM DEM will show systematic errors in areas of higher canopy and smaller precision in areas which are hidden or poorly exposed to the scanning device (Hengl and Reuter, 2008, p.79-80). This opens a possibility to also use the regression-kriging model with auxiliary maps to produce a more realistic error surface (Hengl et al., 2008).

There are three major groups of auxiliary maps of interest to DEM generation:

(1.) Hydrological maps:

- stream line data;
- water stagnation areas (soil-water content images);
- seashore and lakes border lines;

(2.) Land cover maps:

- canopy height;
- Leaf Area Index;
- land cover classes;

(3.) Geomorphological maps:

- surface roughness maps;
- physiographic breaks;
- ridges and terraces;

A lot of topography-connected information can be derived from remote sensing multi- and hyper-spectral images, such as shading-based indices, drainage patterns, ridge-lines, topographic breaks. All these can be derived using automated (pattern recognition) techniques, which can significantly speed up processing for large areas.

Many auxiliary maps will mutually overlap in information and value. Ideally, auxiliary maps used to improve generation of DEMs should be only GIS layers produced independently from the sampled elevations — e.g. remotely sensed images, topographic features, thematic maps etc. Where this is not possible, auxiliary maps can be derived from an existing DEM, provided that this DEM is generated using independent elevation measurements. Care needs to be taken not to employ auxiliary maps which are only indirectly or accidentally connected with the variations in topography. Otherwise unrealistic simulations can be generated, of even poorer quality than if only standard DEM generation techniques are used (Hengl et al., 2008).

2.8 Regression-kriging and sampling optimization algorithms

Understanding the concepts of regression-kriging is not only important to know how to generate maps, but also to know how to prepare a sampling plan and eventually minimize the survey costs. Because the costs of the field survey are usually the biggest part of the survey budget, this issue will become more and more important in the coming years. So far, two main groups of sampling strategies have been commonly utilized for the purpose of environmental mapping (Guttorp, 2003):

- **Regular sampling** — This has the advantage that it systematically covers the area of interest (maximized mean shortest distance), so that the overall prediction variance is usually minimized[24]. The disadvantage of this technique is that it misrepresents distances smaller than the grid size (short range variation).
- **Randomized sampling** — This has the advantage that it represents all distances between the points, which is beneficial for the variogram estimation. The disadvantage is that the spreading of the points in geographic space is lower than in the case of regular sampling, so that the overall precision of the final maps will often be lower.

None of two strategies is universally applicable so that often their combination is recommended: e.g. put half of the points using regular and half using a randomized strategy. Both random and regular sampling strategies belong to the group of design-based sampling. The other big group of sampling designs are the model-based designs. A difference between a design-based sampling (e.g. simple random sampling) and the model-based design is that, in the case of the model-based design, the model is defined and commonly a single optimal design that maximizes/minimizes some criteria can be produced.

In the case of regression-kriging, there are much more possibilities to improve sampling than by using design-based sampling. First, in the case of preparing a sampling design for new survey, the samples can be more objectively located by using some **response surface design** (Hengl et al., 2004b), including the **Latin hypercube sampling** (Minasny and McBratney, 2006). The Latin hypercube sampling will ensure that all points are well-placed in the feature space defined by the environmental factors — these will later be used as predictors — and that the extrapolation in feature space is minimized. Second, once we have collected samples and estimated the regression-kriging model, we can then optimize sampling and derive (1) number of required additional observations and (2) their optimal location in both respective spaces. This leads to a principle of the two-stage[25] model-based sampling (Fig. 2.13).

The **two-stage sampling** is a guarantee of minimization of the survey costs. In the first phase, the surveyors will produce a sampling plan with minimum survey costs — just to have enough points to get a '*rough*' estimate of the regression-kriging model. Once the model is approximated (correlation and variogram model), and depending on the prescribed accuracy (overall prediction variance), the second (additional) sampling plan can be generated. Now we can re-estimate the regression-kriging model and update the predictions so that they fit exactly our prescribed precision requirements. Brus and Heuvelink (2007) tested the use of simulated annealing to produce optimal designs based on the regression-kriging model, and concluded that the resulting sampling plans will lead to hybrid patterns showing spreading in both feature and geographical space. An R package intamap[26] (procedures for automated interpolation) has been recently released that implements such algorithms to run sampling optimization. The interactive version of the intamap package allows users to create either new sampling networks with spatial coverage methods, or to optimally allocate new observations using spatial simulated annealing (see results for the `meuse` case study in Fig. 2.13).

Smarter allocation of the points in the feature and geographic space often proves that equally precise maps could have been produced with much less points than actually collected. This might surprise you, but it has a strong theoretical background. Especially if the predictors are highly correlated with the target variable and if this correlation is close to linear, there is really no need to collect many samples in the study area. In order to produce precise predictions, it would be enough if we spread them around extremes of the feature space and possibly maximized their spreading in the area of interest (Hengl et al., 2004b). Of course, number of sampling points is mainly dictated by our precision requirements, so that more accurate (low overall precision variance) and detailed (fine cell size) maps of environmental variables will often require denser sampling densities.

[24] If ordinary kriging is used to generate predictions.

[25] Ideally, already one iteration of additional sampling should guarantee map of required accuracy/quality. In practice, also the estimation of model will need to be updated with additional predictors, hence more iterations can be anticipated.

[26] `http://intamap.org`

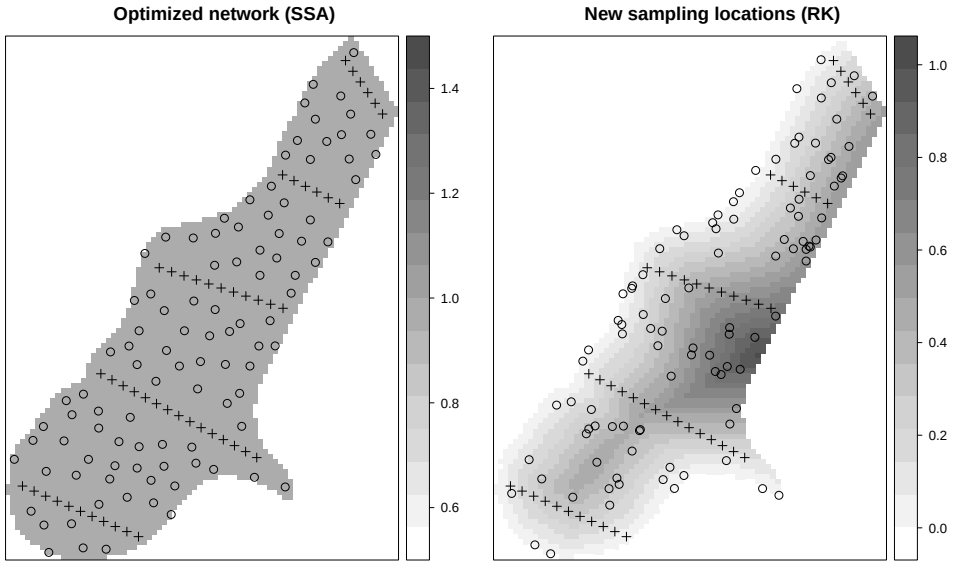

Fig. 2.13: Example of the two-stage model-based sampling: + — 50 first stage samples (transects); o — 100 new samples allocated using the model estimated in the first stage (optimized allocation produced using Spatial Simulated Annealing implemented in the intamapInteractive package). In the case of low correlation with auxiliary maps (left), new sampling design shows have higher spreading in the geographical space; if the correlation with predictors is high (right), then the new sampling design follows the extremes of the features space.

2.9 Fields of application

With the rapid development of remote sensing and geoinformation science, natural resources survey teams are now increasingly creating their products (geoinformation) using ancillary data sources and computer programs — the so-called *direct-to-digital* approach. For example, sampled concentrations of heavy metals can be mapped with higher detail if information about the sources of pollution (distance to industrial areas and traffic or map showing the flooding potential) is used. In the following sections, a short review of the groups of application where regression-kriging has shown its potential is given.

2.9.1 Soil mapping applications

In digital soil mapping, soil variables such as pH, clay content or concentration of a heavy metal, are increasingly mapped using the regression-kriging framework: the deterministic part of variation is dealt with maps of soil forming factors (climatic, relief-based and geological factors) and the residuals are dealt with kriging (McBratney et al., 2003). The same techniques is now used to map categorical variables (Hengl et al., 2007b). A typical soil mapping project based on geostatistics will also be demonstrated in the following chapter of this handbook. This follows the generic framework for spatial prediction set in Hengl et al. (2004a) and applicable also to other environmental and geosciences (Fig. 2.14).

In geomorphometry, auxiliary maps, such as maps of drainage patterns, land cover and remote sensing-based indices, are increasingly used for geostatistical modeling of topography together with point data sets. Auxiliary maps can help explain spatial distribution of errors in DEMs and regression-kriging can be used to generate equiprobable realizations of topography or map the errors in the area of interest (Hengl et al., 2008). Such hybrid geostatistical techniques will be more and more attractive for handling rich LiDAR and radar-based topographic data, both to analyze their inherent geostatistical properties and generate DEMs fit-for-use in various environmental and earth science applications.

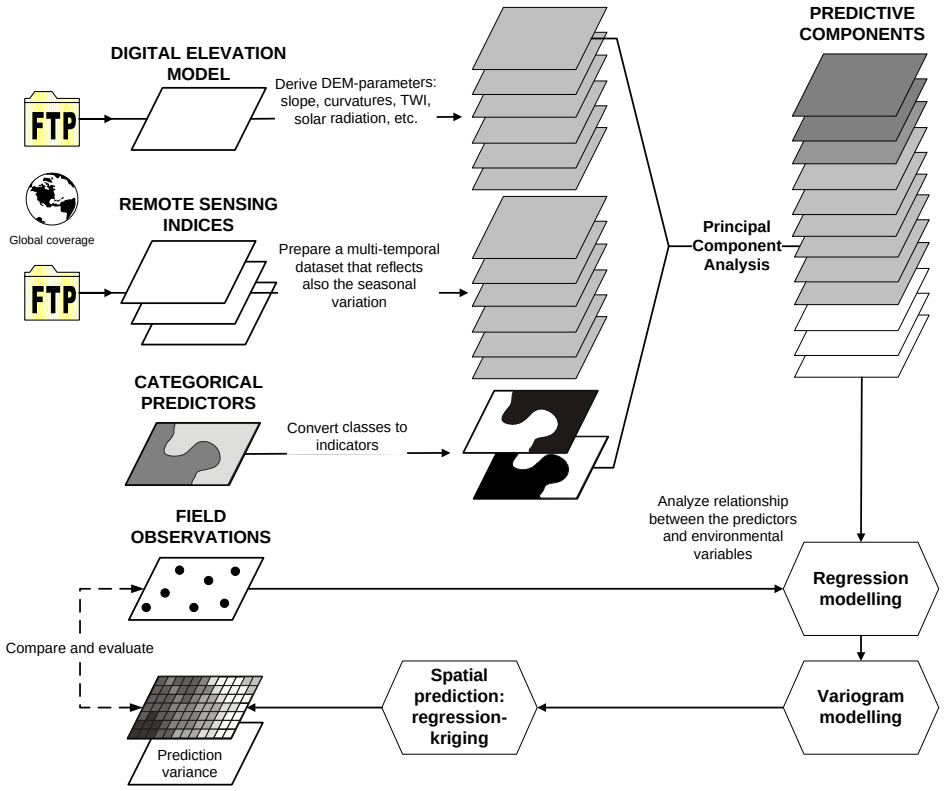

Fig. 2.14: A generic framework for digital soil mapping based on regression-kriging. After Hengl et al. (2007b).

2.9.2 Interpolation of climatic and meteorological data

Regression-kriging of climatic variables, especially the ones derived from DEMs, is now favoured in many climatologic applications (Jarvis and Stuart, 2001; Lloyd, 2005). DEMs are most commonly used to adjust measurements at meteorological stations to local topographic conditions. Other auxiliary predictors used range from distance to sea, meteorological images of land surface temperature, water vapor, short-wave radiation flux, surface albedo, snow Cover, fraction of vegetation cover (see also section 4). In many cases, real deterministic models can be used to make predictions, so that regression-kriging is only used to calibrate the values using the real observations (D'Agostino and Zelenka, 1992, see also Fig. 2.3). The exercise in chapter 11 demonstrates the benefits of using the auxiliary predictors to map climatic variables. In this case the predictors explained almost 90% of variation in the land surface temperatures measured at 152 stations. Such high R-square allows us to *extrapolate* the values much further from the original sampling locations, which would be completely inappropriate to do by using ordinary kriging. The increase of the predictive capabilities using the explanatory information and regression-kriging has been also reported by several participants of the Conference on spatial interpolation in climatology and meteorology (Szalai et al., 2007).

Interpolation of climatic and meteorological data is also interesting because the explanatory (meteorological images) data are today increasingly collected in shorter time intervals so that time-series of images are available and can be used to develop spatio-temporal regression-kriging models. Note also that many meteorological prediction models can generate maps of forecasted conditions in the close-future time, which could then again be calibrated using the actual measurements and RK framework.

2.9.3 Species distribution modeling

Geostatistics is considered to be one of the four spatially-implicit group of techniques suited for species distribution modeling — the other three being: autoregressive models, geographically weighted regression and parameter estimation models (Miller et al., 2007). Type of technique suitable for analysis of species (oc-

currence) records is largely determined by the species' biology. There is a distinct difference between field observation of animal and plant species and measurements of soil or meteorological variables. Especially the observations of animal species asks for high sampling densities in temporal dimension. If the biological species are represented with quantitative composite measures (density, occurrence, biomass, habitat category), such measures are fit for use with standard spatio-temporal geostatistical tools. Some early examples of using geostatistics with the species occurrence records can be found in the work of Legendre and Fortin (1989) and Gotway and Stroup (1997). Kleinschmidt et al. (2005) uses regression-kriging method, based on the Generalized mixed model, to predict the malaria incidence rates in South Africa. Miller (2005) uses a similar principle (predict the regression part, analyze and interpolate residuals, and add them back to predictions) to generate vegetation maps. Miller et al. (2007) further provide a review of predictive vegetation models that incorporate geographical aspect into analysis. Pure interpolation techniques will often outperform niche based models (Bahn and McGill, 2007), although there is no reason not to combine them. Pebesma et al. (2005) demonstrates that geostatistics is fit to be used with spatio-temporal species density records. §8 shows that even occurrence-only records can be successfully analyzed using geostatistics i.e. regression-kriging.

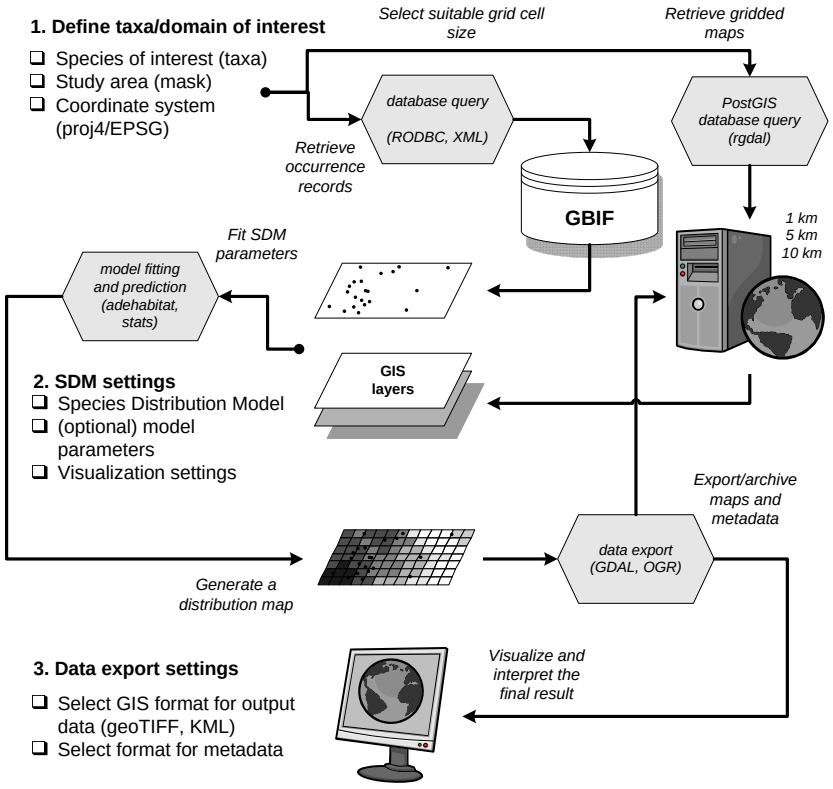

Fig. 2.15: Schematic example of a geo-processing service to automate extraction of species distribution maps using GBIF occurrence records and gridded predictors. The suitable R packages are indicated in brackets.

Fig. 2.15 shows an example of a generic automated data processing scheme to generate distribution maps and similar biodiversity maps using web-data. The occurrence(-only) records can be retrieved from the Global Biodiversity Information Facility[27] (GBIF) Data Portal, then overlaid over number of gridded predictors (possibly stored in a PostGIS database), a species' prediction model can then be fitted, and results exported to some GIS format / KML. Such automated mapping portals are now increasingly being used to generate up-to-date species' distribution maps.

[27] Established in 2001; today the largest international data sharing network for biodiversity.

2.9.4 Downscaling environmental data

Interpolation becomes down-scaling once the grid resolution in more than 50% of the area is finer than it should be for the given sampling density. For example, in soil mapping, one point sample should cover 160 pixels (Hengl, 2006). If we have 100 samples and the size of the area is 10 km^2, then it is valid to map soil variables at resolutions of 25 m (maximum 10 m) or coarser. Note that down-scaling is only valid if we have some auxiliary data (e.g. digital elevation model) which is of finer resolution than the effective grid resolution, and which is highly correlated with the variable of interest.

If the auxiliary predictors are available at finer resolutions than the sampling intensity, regression-kriging can be used to downscale information. Much of recent research in the field of biogeography, for example, has been focusing on the down-scaling techniques (Araújo et al., 2005). Hengl et al. (2008) shows how auxiliary maps can be used to downscale SRTM DEMs from 90 to 30 m resolution. Pebesma et al. (2007) use various auxiliary maps to improve detail of air pollution predictions. For the success of downscaling procedures using regression-kriging, the main issue is how to locate the samples so that extrapolation in the feature space is minimized.

2.10 Final notes about regression-kriging

At the moment, there are not many contra-arguments not to replace the existing traditional soil, vegetation, climatic, geological and similar maps with the maps produced using analytical techniques. Note that this does not mean that we should abandon the traditional concepts of field survey and that surveyors are becoming obsolete. On the contrary, surveyors continue to be needed to prepare and collect the input data and to assess the results of spatial prediction. On the other hand, they are less and less involved in the actual delineation of features or derivation of predictions, which is increasingly the role of the predictive models.

One such linear prediction techniques that is especially promoted in this handbook is regression-kriging (RK). It can be used to interpolate sampled environmental variables (both continuous and categorical) from large point sets. However, in spite of this and other attractive properties of RK, it is not as widely used in geosciences as might be expected. The barriers to widespread routine use of RK in environmental modeling and mapping are as follows. First, the statistical analysis in the case of RK is more sophisticated than for simple mechanistic or kriging techniques. Second, RK is computationally demanding[28] and often cannot be run on standard PCs. The third problem is that many users are confused by the quantity of spatial prediction options, so that they are never sure which one is the most appropriate. In addition, there is a lack of user-friendly GIS environments to run RK. This is because, for many years GIS technologies and geostatistical techniques have been developing independently. Today, a border line between statistical and geographical computing is fading away, in which you will hopefully be more convinced in the remaining chapters of this guide.

2.10.1 Alternatives to RK

The competitors to RK include completely different methods that may fit certain situations better. If the explanatory data is of different origin and reliability, the Bayesian Maximum Entropy approach might be a better alternative (D'Or, 2003). There are also machine learning techniques that combine neural network algorithms and robust prediction techniques (Kanevski et al., 1997). Henderson et al. (2004) used decision trees to predict various soil parameters from large quantity of soil profile data and with the help of land surface and remote sensing attributes. This technique is flexible, optimizes local fits and can be used within a GIS. However, it is statistically suboptimal because it ignores spatial location of points during the derivation of classification trees. The same authors (Henderson et al., 2004, pp.394–396) further reported that, although there is still some spatial correlation in the residuals, it is not clear how to employ it.

Regression-kriging must also be compared with alternative kriging techniques, such as **collocated co-kriging**, which also makes use of the explanatory information. However, collocated co-kriging is developed for situations in which the explanatory information is not spatially exhaustive (Knotters et al., 1995). CK also requires simultaneous modeling of both direct and cross-variograms, which can be time-consuming for large

[28] Why does RK takes so much time? The most enduring computations are connected with derivation of distances from the new point to all sampled points. This can be speed up by setting up a smaller search radius.

number of covariates[29]. In the case where the covariates are available as complete maps, RK will generally be preferred over CK, although CK may in some circumstances give superior results (D'Agostino and Zelenka, 1992; Goovaerts, 1999; Rossiter, 2007). In the case auxiliary point samples of covariates, in addition to auxiliary raster maps, are available, regression-kriging can be combined with co-kriging: first the deterministic part can be dealt with the regression, then the residuals can be interpolated using co-kriging (auxiliary point samples) and added back to the estimated deterministic part of variation.

2.10.2 Limitations of RK

RK have shown a potential to become the most popular mapping technique used by environmental scientists because it is (a) easy to use, and (b) it outperforms plain geostatistical techniques. However, success of RK largely depends on characteristics of the case study i.e. quality of the input data. These are some main consideration one should have in mind when using RK:

(1.) *Data quality*: RK relies completely on the quality of data. If the data comes from different sources and have been sampled using biased or unrepresentative design, the predictions might be even worse than with simple mechanistic prediction techniques. Even a single bad data point can make any regression arbitrarily bad, which affects the RK prediction over the whole area.

(2.) *Under-sampling*: For regression modeling, the multivariate feature space must be well-represented in all dimensions. For variogram modeling, an adequate number of point-pairs must be available at various spacings. Webster and Oliver (2001, p.85) recommend at least 50 and preferably 300 points for variogram estimation. Neter et al. (1996) recommends at least 10 observations per predictor for multiple regression. We strongly recommend using RK only for data sets with more than 50 total observations and at least 10 observations per predictor to prevent over-fitting.

(3.) *Reliable estimation of the covariance/regression model*: The major dissatisfaction of using KED or RK is that both the regression model parameters and covariance function parameters need to be estimated simultaneously. However, in order to estimate coefficients we need to know covariance function of residuals, which can only be estimated after the coefficients (the chicken-egg problem). Here, we have assumed that a single iteration is a satisfactory solution, although someone might also look for other iterative solutions (Kitanidis, 1994). Lark et al. (2005) recently suggested that an iterative Restricted Maximum Likelihood (REML) approach should be used to provide an unbiased estimate of the variogram and regression coefficients. However, this approach is rather demanding for $\gg 10^3$ point data sets because for each iteration, an $n \times n$ matrix is inverted (Minasny and McBratney, 2007).

(4.) *Extrapolation outside the sampled feature space*: If the points do not represent feature space or represent only the central part of it, this will often lead to poor estimation of the model and poor spatial prediction. For this reason, it is important that the points be well spread at the edges of the feature space and that they be symmetrically spread around the center of the feature space (Hengl et al., 2004b). Assessing the extrapolation in feature space is also interesting to allocate additional point samples that can be used to improve the existing prediction models. This also justifies use of multiple predictors to fit the target variable, instead of using only the most significant predictor or first principal component, which if, for example, advocated by the Isatis development team (Bleines et al., 2004).

(5.) *Predictors with uneven relation to the target variable*: Auxiliary maps should have a constant physical relationship with the target variable in all parts of the study area, otherwise artifacts will be produced. An example is a single NDVI as a predictor of topsoil organic matter. If an agricultural field has just been harvested (low NDVI), the prediction map will (incorrectly) show very low organic matter content within the crop field.

(6.) *Intermediate-scale modeling*: RK has not been adapted to fit data locally, with arbitrary neighborhoods for the regression as can be done with kriging with moving window (Walter et al., 2001). Many practitioners would like to adjust the neighborhood to fit their concepts of the scale of processes that are not truly global (across the whole study area) but not completely local either.

[29] Co-kriging requires estimation of $p+1$ variograms, plus $\left[p \cdot (p+1)\right]/2$ cross-variograms, where the p is the number of predictors (Knotters et al., 1995).

(7.) *Data over-fitting problems*: Care needs to be taken when fitting the statistical models — today, complex models and large quantities of predictors can be used so that the model can fit the data almost 100%. But there is a distinction between the goodness of fit and true success of prediction that cannot really be assessed without independent validation (Rykiel, 1996).

If any of these problems occur, RK can give even worse results than even non-statistical, empirical spatial predictors such as inverse distance interpolation or expert systems. The difficulties listed above might also be considered as challenges for the geostatisticians.

2.10.3 Beyond RK

Although the bibliometric research of Zhou et al. (2007) indicates that the field of geostatistics has already reached its peak in 1996–1998, the development of regression-kriging and similar hybrid techniques is certainly not over and the methods will continue to evolve both from theoretical and practical aspect. Gotway Crawford and Young (2008) recognizes four '*hot*' areas of geostatistics that will receive attention in the near future: (1) geostatistics in non-euclidian space (i.e. space that accounts for barriers, streams, disease transmittion vectors etc.); (2) assessment of spatio-temporal support — spatial prediction methods will be increasingly compared at various spatial/temporal scales; users are increasingly doing predictions from point to area support and vice versa; (3) kriging is increasingly used with discrete data and uncertain data (this emphasized the importance of using Bayesian-based models), and (4) geostatistics as a tool of politics.

What you can certainly anticipate in the near future considering regression-kriging connected methods are the following six developments:

- *More sophisticated prediction models*: Typically, regression-kriging is sensitive to blunders in data, local outliers and small size data sets. To avoid such problems, we will experience an evolution of methods that are more generic and more robust to be used to any type of data set. Recently, several authors suggested ways to make more sophisticated, more universally applicable BLUPs (Lark et al., 2005; Minasny and McBratney, 2007; Bárdossy and Li, 2008). We can anticipate a further development of intelligent, iterative data fitting algorithms that can account for problems of local hot-spots, mixed data and poor sampling strategies. This is now one of the major focuses of the intamap project (Pebesma et al., 2009).

- *Local regression-kriging*: As mentioned previously in §2.2, local regression-kriging algorithms are yet to be developed. Integration of the local prediction algorithms (Haas, 1990; Walter et al., 2001) would open many new data analysis possibilities. For example, with local estimation of the regression coefficients and variogram parameters, a user will be able to analyze which predictors are more dominant in different parts of the study area, and how much these parameters vary in space. The output of the interpolation with not be only a map of predictions, but also the maps of (local) regression coefficients, R-square, variogram parameters and similar. Lloyd (2009) recently compared KED (monthly precipitation in UK) based on local variogram models and discovered that it provides more accurate predictions (as judged by cross-validation statistics) than any other 'global' approach.

- *User-friendly sampling optimisation packages*: Although methodologies both to plan new sampling designs, and to optimize additional sampling designs have already been tested and described (Minasny and McBratney, 2006; Brus and Heuvelink, 2007), techniques such as simulated annealing or Latin hypercube sampling are still not used in operational mapping. The recently released `intamapInteractive` package now supports simulated annealing and optimization of sampling designs following the regression-kriging modeling. Development of user-friendly sampling design packages will allow mapping teams to generate (*smart*) sampling schemes at the click of button.

- *Automated interpolation of categorical variables*: So far no tool exists that can automatically generate membership maps given a point data with observed categories (e.g. soil types, land degradation types etc.). A compositional RK algorithm is needed that takes into account relationship between all categories in the legend, and then fits regression models and variogram models for all classes (Hengl et al., 2007b).

- *Intelligent data analysis reports generation*: The next generation of geostatistical packages will be intelligent. It will not only generate predictions and prediction variances, but will also provide interpretation of the fitted models and analysis of the intrinsic properties of the input data sets. This will include detection of possible outliers and hot-spots, robust estimation of the non-linear regression model, assessment

of the quality of the input data sets and final maps. The R package automap, for example, is pointing to this direction.

- *Multi-temporal, multi-variate prediction models*: At the moment, most of the geostatistical mapping projects in environmental sciences focus on mapping a single variable sampled in a short(er) period of time and for a local area of interest. It will not take too long until we will have a global repository of (multi-temporal) predictors (see further section 4.1) and point data sets that could then be interpolated all at once (to employ all possible relationships and cross-correlations). The future data sets will definitively be multi-temporal and multi-variate, and it will certainly ask for more powerful computers and more sophisticated spatio-temporal 3D mapping tools. Consequently, outputs of the spatial prediction models will be animations and multimedia, rather then simple and static 2D maps.

Although we can observe that with the more sophisticated methods (e.g. REML approach), we are able to produce more realistic models, the quality of the output maps depends much more on the quality of input data (Minasny and McBratney, 2007). Hence, we can also anticipate that evolution of technology such as hyperspectral remote sensing and LiDAR will contribute to the field of geostatistical mapping even more than the development of the more sophisticated algorithms.

Finally, we can conclude that an unavoidable trend in the evolution of spatial prediction models will be a **development and use of fully-automated, robust, intelligent mapping systems** (see further §3.4.3). Systems that will be able to detect possible problems in the data, iteratively estimate the most reasonable model parameters, employ all possible explanatory and empirical data, and assist the user in generating the survey reports. Certainly, in the near future, a prediction model will be able to run more analysis with less interaction with user, and offer more information to decision makers. This might overload the inexperience users, so that practical guides even thicker than this one can be anticipated.

Further reading:

★ Banerjee, S. and Carlin, B. and Gelfand, A. 2004. **Hierarchical Modeling and Analysis for Spatial Data**. Chapman & Hall/CRC, Boca Raton, 472 p.

★ Christensen, R. 2001. Best Linear Unbiased Prediction of Spatial Data: Kriging. In: Cristensen, R. **Linear Models for Multivariate, Time Series, and Spatial Data**, Springer, 420 p.

★ Hengl T., Heuvelink G. B. M., Rossiter D. G., 2007. About regression-kriging: from equations to case studies. Computers & Geosciences, 33(10): 1301–1315.

★ Minasny, B., McBratney, A. B., 2007. Spatial prediction of soil properties using EBLUP with Matérn covariance function. Geoderma 140: 324–336.

★ Pebesma, E. J., 1999. **Gstat user's manual**. Department of Physical Geography, Utrecht University, Utrecht, 96 p.

★ Schabenberger, O., Gotway, C. A., 2004. **Statistical methods for spatial data analysis**. Chapman & Hall/CRC, 524 p.

★ Stein, M. L., 1999. **Interpolation of Spatial Data: Some Theory for Kriging**. Series in Statistics. Springer, New York, 247 p.