3. Spatial interpolation models in climatology

3.1 From discrete data to continuous climate surfaces

Spatial interpolation is the procedure of estimating the value of quantities at unsampled sites within the area covered by existing observations. From a discrete set of data, the mathematical spatial models produce continuous climate surfaces (e.g. temperature climatologies). To be correct, they produce grids (often called rasters in geostatistics) that depend on the resolution of a digital elevation model (the so-called *DEM*), e.g., approximately 30-arc-second for each grid cell for the U.S. Geological Survey *DEM (USGS website)*, the *DEM* used by our research group.

Geostatistical mapping can be defined as the analytical production of maps by using field observations, explanatory information and a computer program, which calculates values at locations of interest (i.e. a study area).

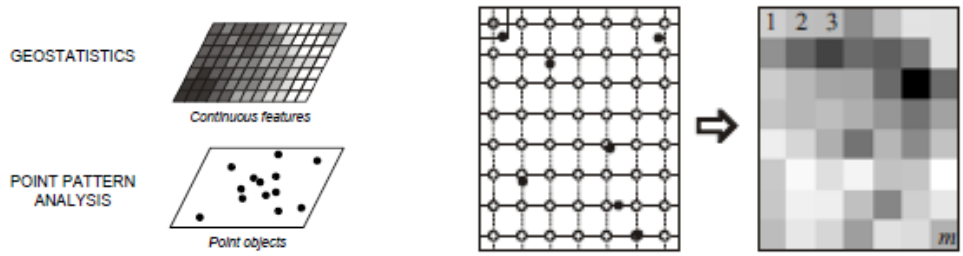

Fig.9 From point data to continuous field based on digital model (Cressie, 1993; Hengl, 2009)

Spatialization models can be classified into:

- Mechanical Deterministic models: Thiessen polygons, inverse distance interpolation, regression on coordinates, natural neighbours, splines and so on.
- Linear Statistical Probability models: kriging, environmental correlation models, Baesyan-based models, hybrid models and so on.
- Expert-Based models: purely data-driven models, *PRISM*, neural networks and so on.

The output of such global or moving window models can be smoothed or exact, gridded or in polygons, depending on the application of the model itself.

Climate spatialization models are usually univariate: a dependent variable (e.g. mean temperature, precipitation, air pressure and so on) and a set of independent variables that can be studied in subsequent steps or in a single model. The independent variables are also known as predictors and it is very important to have a clear idea about connection or hierarchy of predictors (*Hengl, 2009)*. Independent variables can be large scale variables (latitude, longitude, distance from the coast, land cover classes, slope steepness, aspect, solar radiation), vertical parameters (elevation), local variables (proximity to lakes, cold air pools, urban heat island) and they contribute, besides data availability and density, to choosing the statistical model *(Jarvis et al, 2001a; Jarvis et al., 2001b; Daly, 2006)*.

It must be underlined that the selection of modelling method is as important as data quality and as the choice of independent variables *(Jarvis et al., 2001a).*

A spatialization model must be validated using statistical parameters that quantify errors (Mean Absolute Error, Root Mean Square Error, Bias, and so on) and statistical techniques as cross validation, boot-strapping, and so on.

In this chapter, we describe the most used spatialization models to predict climate surfaces, citing some examples where the models were applied in climate research and briefly discussing about comparisons between models. In the last section, we list the validation methodologies.

3.2 Kriging

3.2.1 Definitions and assumptions

Based on mathematical calculations of George Matheron *(Matheron, 1969)*, following Daniel Krige's master thesis *(Krige, 1951)*, kriging is a stochastic spatial interpolation method which uses locally weighted averages. The weights are derived from variograms and kriging treats statistically random variations in space for the regionalized variable *(Jarvis et al., 2001a; Chapman et al., 2003)*. Such weights depend on reciprocal distance between grid points *(Gylfadottir, 2003)*.

Kriging is performed under some basic assumptions: data should have stationarity, unbiasedness and a normal statistic distribution, global trends must be removed from data before kriging and the predictors (i.e. the independent variables) must be known at every

grid point of the raster where the predictands (i.e. the dependent variables) are calculated *(Hengl et al., 2003)*. If no trend is present in the data, kriging gives a *BLUE* (Best Linear Unbiased Estimation) interpolation *(Ashraf et al., 1997)*.

3.2.2 Kriging models: variants, peculiarities, and limits

Here we list the most used variants of kriging (*Hengl et al., 2003; Hengl et al. 2009; Stein, 1999*):

- Simple Kriging (*SK*): the expectation value is zero everywhere, the covariance is known
- Ordinary Kriging (*OK*): the expectation value is unknown but constant, the variogram is known
- Co-kriging (*CK*): a multivariate kriging that uses a regionalized variable theory based on inter-variable correlations, it uses cross-semi-variograms
- Universal Kriging (*UnK*): the expectation value is described with a general linear model
- IRFk-kriging (*IRFK*): the expectation value is described with a polynomial model
- Indicator kriging (*IK*): it uses indicator functions to estimate transition probabilities
- Multiple indicator kriging (*MIK*): an indicator kriging that uses a family of indicators
- Disjunctive kriging (*DJK*): a non-linear generalisation of kriging
- Lognormal kriging (*LNK*): it interpolates positive data by means of logarithms
- Kriging with external drift (*KED*): a deterministic variable is first used to de-trend the data, then ordinary kriging is performed
- Regression Kriging (*RK*): drift components and residuals are treated separately and then summed, the former are studied with regression analysis, the latter with ordinary kriging

Kriging can use only station data, as in simple kriging, block kriging, ordinary kriging, or it can use also topographic information, as in universal kriging or co-kriging *(Vicente-Serrano et al., 2003; Isaaks et al., 1989)*.

If all the assumptions are verified, *OK* or *SK* can be used, but unfortunately climate data is usually biased and stationarity is rarely verified.

To overcome this problem, deterministic components are first de-trended, e.g. with a multiple linear regression of temperature versus longitude, latitude, elevation, distance from the coast, then kriging is performed over residuals. This technique is called residual kriging (*ResK*) *(Tveito, 2002)*.

UK, KED, RK, ResK are "hybrid" non-stationarity geo-statistical methods that are called "kriging after de-trending" methods and such methods are formally very similar *(Wackernagel, 2003*; *Goovaerts, 2000)*. They should be used when the auxiliary information is available for every grid point and if a trend is present in the data. These three methods provide quite identical results if the input parameters are the same.

KED calculates weights by expanding the covariance matrix with auxiliary variables, which ensure that universality is included in the kriging system. *RK* solves the drift model coefficients with regressions then it uses *OK* to interpolate residuals and it added them back to get final results *(Hengl et al., 2003)*. *ResK* is a more general *RK* method that solves the drift components with regressions or other models. *UnK* is a kriging with a changing mean where the trend is modelled as function of coordinates and it should be used only in this case *(Matheron, 1969; Wackernagel, 2003)*. It can be demonstrated that *KED* and *RK* are based on the same mathematical equations *(Hengl et al., 2003)*.

Furthermore, *RK* turns into *SK* or *OK* if predictors and predictands are significantly not correlated, whilst it turns into pure regressions if predictors and predictands are significantly correlated. Thus, *MLR* (Multiple linear regression) models and first order kriging models are special cases of *RK (Hengl, 2009)*.

On the other hand, if the kriging variable was not sampled enough, *CK* or collocated co-kriging (*CCK*) provides better estimations *(Ashraf et al. 1997; Knotters et al., 1995)* and they are based on fewer spatial variables than *SK* or *OK (Goovaerts et al., 2000)*.

If the datasets are very large, the so-called sequential kriging (*SeqK*) can be used. It is based on sequentially data selection of subsets where simple or ordinary kriging are applied *(Vargas-Guzman et al., 1999)*. *SeqK* includes local-regression kriging (*LRK*) models or moving window kriging models that were used in recent analysis *(Lloyd, 2010*; *Walter et al., 2001)*.

The kriging technique has some limitations. Kriging relies completely on data quality. If the dataset is small, there could be over-fitting and under-sampling. In *KED* or *RK,* the covariance and regression calculations are performed simultaneously. Extrapolation outside the sampled feature area could produce biased results, predictors with uneven relation to the target variable must be avoided, intermediate-scale modelling can be difficult and there is still a lack of *GIS* user-friendly packages dedicated to kriging

(Hengl, 2009). Moreover, it should be noticed that kriging variograms are calculated using distance as parameter and this can be reductive in areas with a complex orography as Italy.

3.2.3 Examples of kriging models: *OK* and *RK*

• Data with no intrinsic trend: ordinary kriging equations

This paragraph is based on *Hengl (2009)*.

If the trend function is constant, the variogram is constant and the target variable follows approximately a normal distribution, the *OK* method can be used *(Hengl, 2009)*.

If Z is the "target variable", $z(s_1)$, $z(s_2)$,. . . , $z(s_n)$ is a set of observations and $\underline{s}_i = (x_i, y_i)$ are geographical locations, the predictions are based on the model:

$$Z(\underline{s}) = \mu + \varepsilon'(\underline{s}) \quad \textbf{(1)}$$

Where μ is the constant stationary function (global mean) and $\varepsilon'(\underline{s})$ is the spatially correlated stochastic part of variation.

Thus, predictions are made as:

$$\hat{z}_{OK}(\underline{s}_0) = \sum_{i=1}^{n} w_i(\underline{s}_0) \cdot z(\underline{s}_i) = \lambda_0^T \cdot \underline{z} \quad \textbf{(2)}$$

Where λ_0 is the vector of kriging weights, $\underline{z}$ is the vector of n observations at primary locations.

In order to estimate weights in an objective way, a variogram is used A variogram is defined as a function describing the degree of spatial dependence of the spatial random field or stochastic process $Z(\underline{s})$ *(Wackernagel, 2003)*.

The squared differences between the neighbouring values are called semi-variances:

$$\gamma(\underline{h}) = \frac{1}{2} E\left[(z(\underline{s}_i) - z(\underline{s}_i + \underline{h}))^2\right] \quad \textbf{(3)}$$

where $z(\underline{s}_i)$ is the value of the target variable at some sampled location, $z(\underline{s}_i+\underline{h})$ is the value of the neighbour at distance $\underline{s}_i+\underline{h}$.

If we plot all semi-variances versus their separation distance and then we average the values for a standard distance ("lag"), we get an experimental variogram.

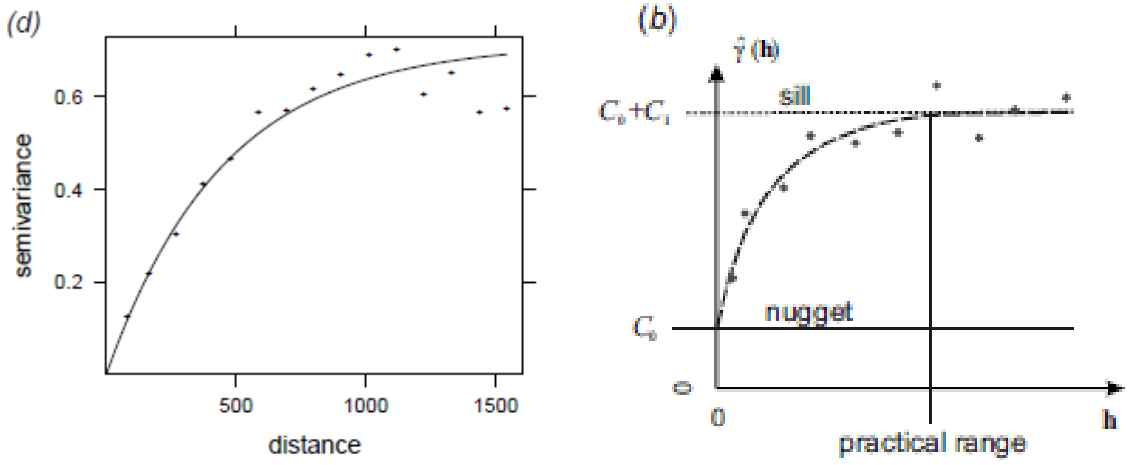

Fig.10 Left: example of variogram: semi-variances vs. pair distances aggregated to standard lags. Right: sill variation and nugget value in semivariograms to find the practical range. (Hengl, 2009)

The experimental variogram is fitted with a theoretical model, e.g. linear, spherical, exponential, circular, Gaussian, Belle, power and so on (*Isaaks et al., 1989; Goovaerts, 2000).*

The *OK* weights are solved by multiplying the covariances:

$$\lambda_0 = \underline{C} \cdot \underline{c}_0 \quad ; \quad C(|\underline{h}| = 0) = C_0 + C_1 \quad \textbf{(4)}$$

where *C* is the covariance matrix for *n x n* observations and $\underline{c}_0$ is the vector of covariances at a new location. $\underline{C}$ is *(n+1) x (n+1)* matrix if used to derive kriging weights.

We have:

$$\begin{bmatrix} C(\underline{s}_1,\underline{s}_1) & \dots & C(\underline{s}_1,\underline{s}_n) & 1 \\ \dots & \dots & \dots. & \dots \\ C(\underline{s}_n,\underline{s}_1) & \dots & C(\underline{s}_n,\underline{s}_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} * \begin{bmatrix} C(\underline{s}_0,\underline{s}_1) \\ \dots \\ C(\underline{s}_0,\underline{s}_n) \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(\underline{s}_0) \\ \dots \\ w_n(\underline{s}_0) \\ \varphi \end{bmatrix} \quad \textbf{(5)}$$

Where one extra row and column are used to ensure that the sum of weights is normalized and φ is the Lagrange multiplier.

The OK variance is defined as the weighted average of covariances from the new point ($\underline{s}_0$) to all calibration points ($\underline{s}_1$, …, $\underline{s}_n$) plus the Lagrange multiplier (*Webster et al., 2001*):

$$\hat{\sigma}^2_{OK}(\underline{s}_0) = C_0 + C_1 - \sum_{i=1}^{n} w_i(\underline{s}_0) \cdot C(\underline{s}_0,\underline{s}_i) + \varphi \quad \textbf{(6)}$$

• Data with an intrinsic trend: residual kriging equations

This paragraph is based on *Hengl (2009)*.

RK is a kriging model that can be seen as a *BLUP* (Best Linear Unbiased Prediction) model: *OK*, environmental correlation, *ID* (Inverse distance), *NN* (nearest neighbour) models are special cases of *RK (Hengl, 2009)*.

If data are not stationary, a value of target variable at some location can be seen as the sum of deterministic and stochastic components that can be modelled separately *(Jarvis et al., 2001a; Matheron, 1969)*:

$$Z(\underline{s}) = m(\underline{s}) + \varepsilon'(\underline{s}) + \varepsilon'' \quad \textbf{(7)}$$

Thus, we obtain:

$$\hat{z}(\underline{s}_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\underline{s}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\underline{s}_i) \quad \textbf{(8)}$$

Where the first term in the sum is the deterministic part and the second is the interpolated residual. The regression coefficients $\hat{\beta}_k$ can be estimated by fitting methods as *OLS* (Ordinary Least Squares) or *GLS* (Generalized Least Squares), λ_i are kriging weights determined by the spatial dependence structure of the residuals, $e(\underline{s}_i)$ is the residual at $\underline{s}_i$.

When the regression coefficients ($\hat{\beta}_{GLS}$) and thus the deterministic part are determined, the residual is interpolated with kriging and added to estimated trend.

We can write, in matrix notation, the prediction values as *(Christensen, 2001)*:

$$\hat{z}_{RK}(\underline{s}_0) = \underline{q}_0^{\,T} \cdot \hat{\beta}_{GLS} + \lambda_0^T \cdot (\underline{z} - \underline{q} \cdot \hat{\beta}_{GLS}) \qquad \textbf{(9)}$$

Where the predictions are made at location $\underline{s}_0$, $\underline{q}_0$ is the *(p + 1)* vector of predictors and λ_0 is the vector of *n* kriging weights used to interpolate the residuals, $\underline{q}$ is a matrix of predictors of the sampling locations, $\underline{z}$ is the vector of measured values of the target values.

This *BLUP* model has a *RK* variance of:

$$\hat{\sigma}^2_{RK}(\underline{s}_0) = C_0 + C_1 - \underline{c}_0^T \cdot \underline{C}^{-1} \cdot \underline{c}_0 + (\underline{q}_0 - q^T \cdot \underline{C}^{-1} \cdot \underline{c}_0)^T \cdot (\underline{q}^T \cdot \underline{C}^{-1} \cdot \underline{q})^{-1} \cdot (\underline{q}_0 - \underline{q}^T \cdot \underline{C}^{-1} \cdot \underline{c}_0) \qquad \textbf{(10)}$$

Where $C_0 + C_1$ is the sill variation, $\underline{c}_0$ is the vector of covariances of residuals at the unvisited location and $\underline{C}$ is the covariance matrix of the residuals. If the residuals are not spatially correlated, the *RK* converges to *MLR* because $\underline{C} = \underline{I}$.

Further details on *OK, RK, UK* and *KED* can be found in *Hengl (2009), Hengl et al. (2003), Cressie et al. (1993*), on *SK* and *SeqK* in *Vargas-Gùzman et al. (1999*), on variograms and semi-variograms in *Jarvis et al. (2001a), Vicente-Serrano et al. (2003), Ashraf et al. (1997)*.

3.2.4 Spatial interpolation using kriging models in climate studies

Many recent examples of climatologies, or other applications realized using kriging methods, can be found in literature. The most used kriging types are *OK* and *ResK*, but rarely *CK*. Just to cite some cases, *OK* was used for rainfall climatologies for Australia by *Jeffrey et al. (2001)*, for temperature and precipitation for Lake district (Minnesota, USA) area by *Holdaway (1996)*, for temperature and growing season variables for Southern Qilian mountains (China) by *Chuayan et al. (2005*), for daily temperature for southern France by *Courault et al. (1999)*; *Res-K* as *MLR plus OK* was used for Nordic climate maps for Scandinavia by *Tveito et al. (2001), Tveito et al. (2005)*, for winter temperatures for Norway by *Tveito et al. (2002)*, for temperature and precipitation for the Mediterranean area by *Agnew wt al. (2000)*, for multi-variable climatologies for Poland by *Ustrnul et al. (2003), Ustrnul et al. (2005), Bialobrzeska et al. (2009)*, for temperature for Iceland by *Gylfadòttir, (2003)*; *Res-K* as *KED* or *UK* was used for temperature for Spain mountainous areas by *Benavides et al. (2007)* and for temperature for Scotland by *Hudson et al. (1994)*.

Kriging is also used in weather forecasting models, e.g. for Finland by *Venalainen et al. (2002)* or for Hungary (*MISH* models) by *Bihari et al. (2009)*.

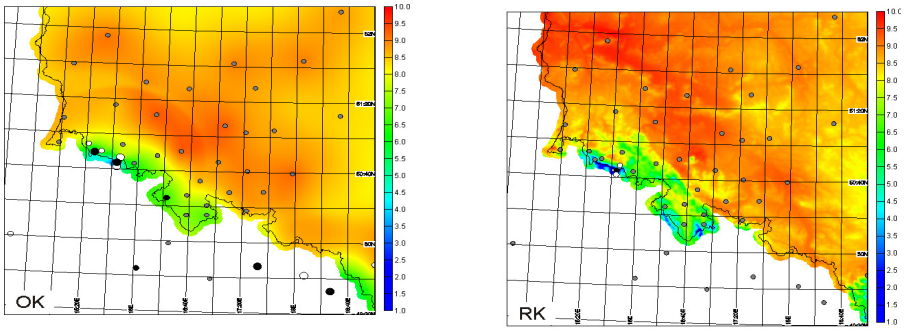

Fig.11 Yearly 61-90 mean temperature climatologies for Poland with OK and RK (Bialobrzeska et al., 2009)

3.2.5 Comparisons between kriging and similar spatialization methods

The kriging method is a widespread spatialization model in climate studies, but many authors disagree about its supremacy in constructing climatologies.

Kriging usually leads to better results and lower errors if compared with simpler models, such as Thiessen polygons (or *NN*) (e.g. temperature and precipitation

climatologies for the Ecuadorian Andes, *KED*, *Butyaert et al., 2006)*, *IDW* (Inverse Distance Weighting) or *ISDW* (Inverse Squared Distance Weighting) (e.g. *Ashraf et al., 1997*).

If compared with regression methods, the question is open. *MLR* or stepwise *LR* (Linear Regression) models give better results for temperature and precipitation in Spanish mountains *(Benavides et al., 2001)*, for winter precipitation and temperature in New Mexico *(Brown et al., 2002)*, for temperature in Qilian area (but in this case *OK* performs better in growing season, *(Chuanyan et al., 2005)*, for monthly rainfall in southern Spain *(Vicente-Serrano et al., 2003)*. Whilst *KED* kriging models show more accurate spatial climatologies for climate maps of Arizona *(Skirvin et al., 2003)* and so on.

TPS (Thin Plate Splines) usually outperform kriging because in real datasets the normal distribution and the stationarity requests are not fulfilled and also because they are computationally handled and modified more easily *(Hartkamp et al., 1999)*. Another example where splines give better results are daily minimum and maximum temperature for the United Kingdom *(Jarvis et al., 2001b)*. On the other hand, for *Chuayan et al. (2005)*, *OK* is better than splines.

Res-K models usually perform better than simple kriging models: e.g. *UnK* is statistically better than *OK* for *Benavides et al. (2007)*, *RK* is better than *OK* and *CK* for *Ustrnul et al. (2005)*, *RK* is better than *OK* for solar radiation maps for Andalusia *(Alsamamra et al., 2009)*.

3.3 Regression models

3.3.1 Definitions and assumptions

In statistics, regression analysis include any techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables *(Lindley, 1987)*.

Regression analysis include a family of functions called *GLMs* (Generalized Linear Models), which assume a linear relationship between the inputs and the outputs *(Hengl, 2009)*. The output from the model fitting is a set of regression coefficients.

If, in the area of interest, there are co-variables significantly correlated with the target variable and if the point values are not totally auto-correlated, regression models between the target variable versus the covariates can be used. Only the deterministic part of variation is taken into account *(Hengl, 2009)*.

In order to use regression models, we have to fulfil some basic assumptions. The sample must be representative of the "population" for the period under investigation. The error must be a random variable and it is assumed to be zero if averaged on the explanatory variables (i.e. such variables must be error free). The predictors should be linearly independent, the errors must be uncorrelated and the variance must be approximately constant across the observations. Furthermore, this technique can be used only if the explanatory variables are known in every grid point; that is the raster surfaces must be available for the given predictors.

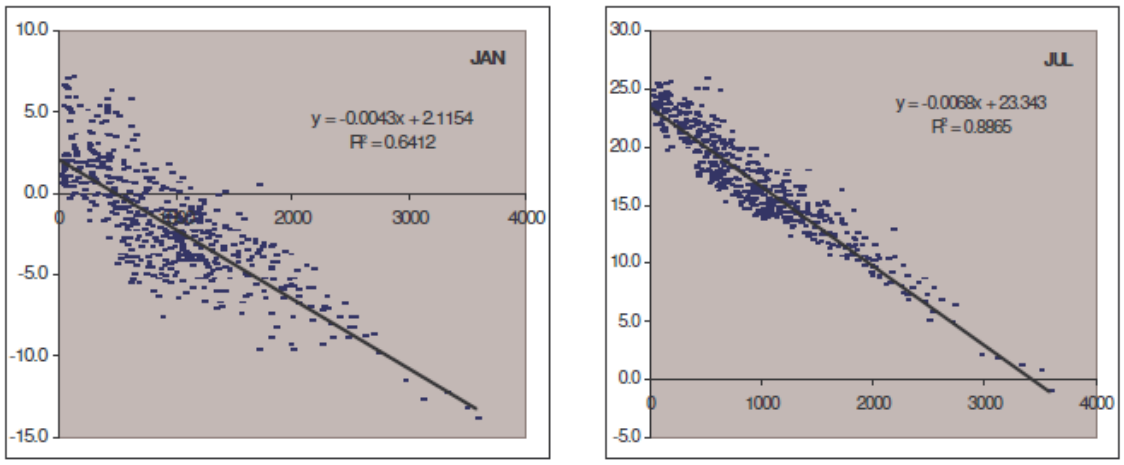

Fig.12 Example of a regression analysis with one independent variable: temperature vs. elevation for 1961-90 in Italy, January (left) and July (right) (Spinoni et al., 2009)

3.3.2 Regression models: variants, limits, and explanatory variables

Regression models belong to "environmental correlation" models. Using factors of climate, topography, geology and so on, regression models deal with the deterministic part of variation.

The target variable is the dependent variable and the explanatory variables are the independent variables. The main advantage of regression models in climate applications is

that the strong physical meaning and the explanatory variables can be chosen by a priori empirical knowledge about the variation of the target features in the area of interest *(Hengl, 2009)*. The most used explanatory variables are geographic variables as longitude, latitude, distance from the coast, topographic variables as elevation, slope, aspect, climatic proxies as land cover, other physical variables as cloudiness, solar radiation, relative humidity, local binary parameters (yes or no conditions) as cold air pools drainage, lakeshores, urban heat islands and so on. The recent availability of such environmental predictors makes this regression approach very common and accurate *(Pebesma, 2006; Hengl et al., 2003)*.

There are many types of regression models in climate research *(Hengl, 2009; Kurtzman et al., 1999; Stahl et al., 2006; Dodson et al., 1997; Goodale et al, 1998; Srivastava, 1998; Bolstad, 2007*):

- Multiple Linear Regression (*MLR*): it models the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data;
- Stepwise Linear Regression (*SLR*): for each step, only one predictor is regressed; there are two different methods, i.e. forward selection of predictors (starting with no variables in the model, trying out the variables one by one and including them if they improve the variance estimation) and backward elimination (or forced entry, starting with all candidate variables and testing them one by one for statistical significance, deleting the not significant ones);
- Geographic Weighted Regression (*GWR*): it fits the regression models by using a moving window algorithm, thus producing maps of regression coefficients that help to analyze how much the model is dependent on the location;
- Local Lapse Rate Regression (*LLR*): based on target variable versus elevation lapse rates;
- Non Linear Regression (*NLR*): mixed models that use explanatory variables as logarithmic, squared, exponential and so on variables as predictors;
- Multivariate Regression (*MVR*): rarely used in climate research, many target variables are studied simultaneously versus common explanatory variables, the Baesian multivariate linear regression (*BMLR*) as an example.

Each of these methods produces a regressed fit and the so-called residuals. In many cases the residuals, i.e. the stochastic part of the variation, are further interpolated using splines, kriging or inverse distance methods *(Ninyerola et al., 2007)*. This following interpolation leads to reducing the errors, but if the dataset is not dense enough, it can

introduce spatial patterns that are only based on statistical improvement of the explained variance and have no physical meaning. If the residuals are plotted considering only features that can be explained by physical reasons, the combination of *MLR* models and residual interpolation methods leads to very accurate climatologies. This is one of the reasons (see Chapter 4) that convinced us to choose an *MLR* plus a weighted residual interpolation for the temperature 1961-1990 models for Italy.

The *RK* (Regression Kriging) model is based on the same assumptions of the *MLR plus OK* of residuals technique *(Hengl et al., 2003)*.

The main limitation of regression models consists in the non-reproducibility in some cases of such models. In fact, the regression coefficients are often local or regional and they could not be applied to other areas, seasons or scales *(Hengl, 2009)*.

3.3.3 Example of regression model: general *MLR*

This paragraph follows *Hengl (2009)*. If all the assumptions in chapter 3.3.1 are fulfilled, we can obtain predictions of the target variable just by studying the deterministic part of the variation:

$$Z(\underline{s}) = f\left(q_k(\underline{s})\right) + \varepsilon \qquad \textbf{(11)}$$

Where q_k are the auxiliary predictors (i.e. the explanatory variables).

In the *MLR* model *(Draper et al., 1998)*, the predictions are obtained by weighted averaging the predictors:

$$\hat{z}_{OLS}(\underline{s}_0) = \hat{b}_0 + \hat{b}_1 \cdot q_1(\underline{s}_0) + \ldots + \hat{b}_p \cdot q_p(\underline{s}_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\underline{s}_0) = \hat{\beta}_k^{\,T} \cdot \underline{q} \qquad q_0(\underline{s}_0) \equiv 1 \qquad \textbf{(12)}$$

Where $q_k(\underline{s}_0)$ are the values of the explanatory variables at the target location, p is the number of explanatory variables, $\hat{\beta}_k$ are the regression coefficients solved, for example, using the *OLS* (Ordinary Least Squares) method, $\underline{q}$ is the *(n x p + 1)* matrix of the predictors.

In a climate *MLR* model, e.g., we can set Z = mean temperature, q_1 = longitude, q_2 = latitude, q_3 = elevation, q_4 = distance from the coast, $\underline{s}_0$ = (longitude, latitude).

The prediction error of a *MLR* model is *(Neter et al., 1996)*:

$$\hat{\sigma}^2_{OLS}(\underline{s}_0) = MSE \cdot \left[1 + \underline{q}_0^T \cdot (q^T \cdot \underline{q})^{-1} \cdot \underline{q}_0\right] \quad \textbf{(13)}$$

Where $\underline{q}_0$ is the vector of predictors at a new and unvisited location and *MSE* is the Mean Squared Error around the regression line, defined as:

$$MSE = \frac{\sum_{i=1}^{n} \left[z(\underline{s}_i) - \hat{z}(\underline{s}_i)\right]^2}{n-1} \quad \textbf{(14)}$$

The prediction error depends on the *MSE*, on the spreading of points in the feature space, on the 'deviation' of the new observation from the centre of the feature space. So, if the model is linear, we can increase the prediction variance by decreasing the spreading of the points in feature space.

The sum of squares of residuals (*SSE*) can be used to determine the adjusted coefficient of multiple determination (R^2_a), that is an indicator of the goodness of fit:

$$R^2_a = 1 - \left(\frac{n-1}{n-p}\right) \cdot (1 - R^2) \quad \textbf{(15)}$$

Where p is the number of explanatory variable, n is the number of data observations, R^2 is the explained variance.

If $R^2_a > 0.85$, the regression model is said to be very satisfactory in climate science. However, the estimation of R^2_a or simply R^2 is often not enough to validate a regression model. It is better to perform a graphical qualitative and sometimes even a quantitative analysis of residuals, in order to see if some areas are uniformly biased and trends were not removed (see Chapter 3.8).

3.3.4 Spatial interpolation using regression models in climate studies

Many applications of various types of regression models can be found in literature. Pure *MLR* models were used for climate maps for USA by *Ollinger et al. (1995)*, for temperature versus longitude, latitude, solar radiation, continentality index and cloudiness for Cataluña (Spain) by *Ninyerola et al. (2000)*, for temperature versus longitude, latitude, elevation and distance from the coast for Arctic Canada by *Atkinson et al. (2002)*, as local lapse rates regressions by *Dodson et al. (1997)*. Polynomial regressions were used for temperature and growing season index for Ireland by *Goodale et al. (1998)*, for temperature for the *UnK* by *Lennon et al. (1995*), for temperature versus longitude, latitude, elevation, distance from the coast plus local regressions for *GAR* (Greater Alpine Area) by *Hiebl et al. (2009)* (see fig. 13), for many climatic variables for Europe by *Bissolli et al. (2009)*.

Stepwise regression models were used, e.g., for temperature for Israel by *Kurtzman et al. (1999)* and for urban daily temperatures in Goteborg (Sweden) by *Eliasson et al. (2003)*. Whilst Multiple Weighted Linear Regression (*MWLR*) was used for snow cover interpolation for Czech Republic by *Stepanek et al. (2009)* and Geographic Weighted Regression (*GWR*) was used by *Szimanovski et al. (2009)* for urban temperatures in Wrocaw (Poland). More frequently, *MLR* models are used in combination with kriging of residuals, as for temperatures for Iceland by *Bjornsson et al. (2003)*, with *IDW* and splines of residuals, as for temperature for Spain by *Ninyerola et al. (2007)*, with optimal interpolation of residuals, as for the Croatian Atlas by *Percec-Tadic et al. (2002)*.

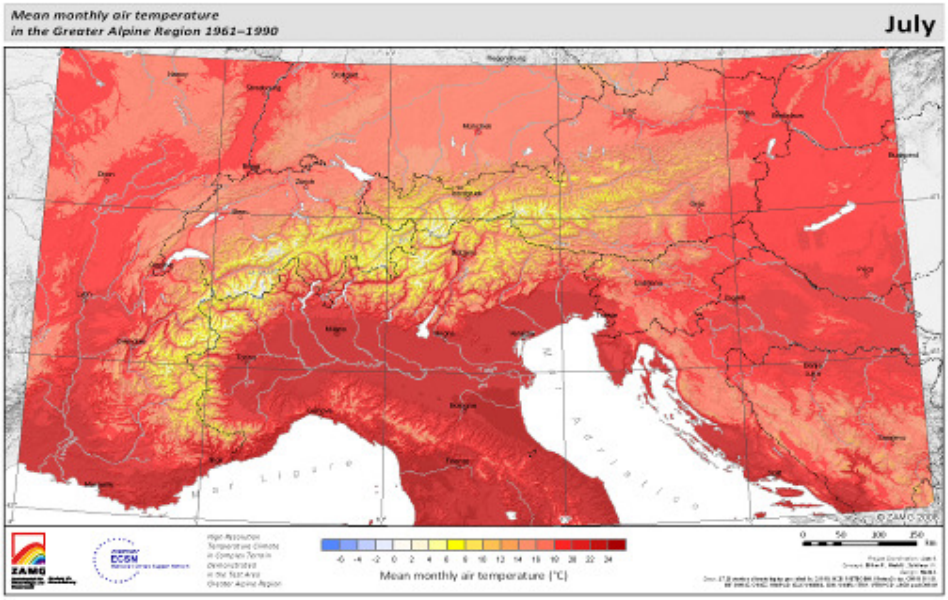

Fig.13 1961-90 mean July temp. climatology for GAR by MLR plus local improvements (Hiebl et al., 2009)

3.3.5 Comparisons between regression models and similar methods

In areas where the orography is very complex and where many water bodies and narrow valleys are present, regression models plus local improvement based on local climate a-priori knowledge are probably the best choice for the construction of temperature or precipitation climatologies.

As seen in Chapter 3.2, this approach is similar to residual kriging. See Chapter 3.2.2 for further details and Chapter 3.2.5 for comparisons between kriging and regression models. *Ninyerola et al. (2002)* and *Ninyerola et al. (2007)* state that *MLR* models are the most accurate models dealing with mountainous regions, and they suggest combining *MLR* and *IDW* or splines as in *Lennon et al. (1995*) for residuals interpolations.

MLR usually performs better than *NN, IDS* or *IDW* models, but in some cases these methods give similar results (*Goodale et al., 1998)* or *GIDS* outperforms *MLR* (*Stahl et al., 2006)*.

Splines and *MLR* models are rarely compared because they are very different methods *(Kurtzman et al., 1999)*.

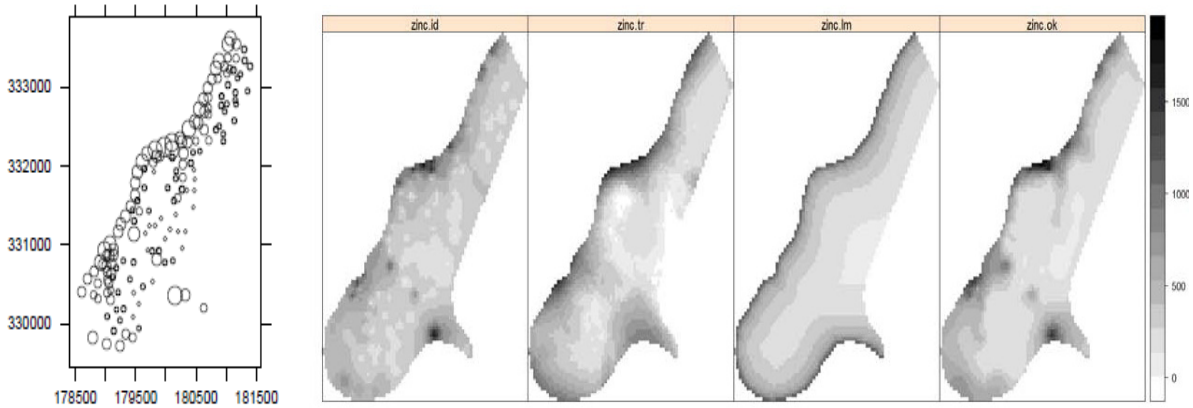

Fig.14 Left: sample locations and measured values of a target variable. Right: spatialization of the target variable with ID, TS (Trend Surface), LR (vs. distance from the river), OK methods (Hengl, 2009)

3.4 Inverse distance models

3.4.1 Definitions and assumptions

Inverse distance weighting (*IDW*) is an interpolation method that assigns values to unknown points by using values from a dataset of known points. The value at each

unknown point is a weighted sum of the values at the known surrounding points *(Shepard, 1968)*. *IDW* is based on the assumption that the interpolating surface should be influenced most by nearby points and least by the more distant points. Such an interpolating surface is a weighted average of the scatter points and the weight assigned to each scatter point diminishes as the distance from the interpolation point to the scatter point increases. Thus, the weighting factor is the distance between points but, to the resulting error in the interpolated estimates, the search parameter (which decides how many points are included in the weighting procedure) is less critical than the power function *(Hodgson et al., 1993)*.

3.4.2 *IDW* models: variants, peculiarities, and limits

Inverse distance weighting models are the evolution of nearest neighbours (*NN*) models. Their principal advantage consists in the fast computational running of this technique because of the simplicity in the formalism. The main limit consists in the limited possibilities of *IDW*. It rarely includes explanatory variables other than distance between points or, in some cases, elevation or longitude and latitude as external drifts. Nevertheless, *IDW* models are a powerful tool when residuals that must be interpolated for weights can be adjusted even following the a-priori knowledge of climate researchers.

We can list many variants of *IDW* models *(Shepard, 1968; Curtis et al., 1999; Nalder et al., 1998; Hofstra et al., 2008; Pan et al., 2004; Stahl et al., 2006)*:

- Inverse Distance Weighting (*IDW*): the basic algorithm is the Shepard's method; usually in literature the acronym *IDW* means that weights are based on the inverse of the distance, i.e. the power is equal to 1; an alternative is the modified Shepard method that calculates interpolated value using only the nearest neighbours within *R*-sphere instead of full sample; another choice is the Lukaszyk-Karmowki metric's modification of the Shepard model;
- Inverse Squared Distance Weighting (ISDW): or simply *IDS* or *ISD* weights are based on the squared inverse of the distance;
- Angular Distance Weighted (*ADW*): distance is weighted with angular functions;
- Gradient plus Inverse Distance Squared (*GIDS*): *MLR* followed by *IDS* interpolation of the residuals;

- Smart Distance Searching (*SDS*): includes latitude, longitude and elevation de-trending before IDW interpolation, similar to *GIDS* but formally not identical.

Our temperature model for Italy is a *MLR* model, plus local improvements, plus *IDW* with Gaussian weights of the residuals, thus it can be considered a further development of the *GIDS* model, that should be used only if the data density is high *(Price et al., 2000)*.

3.4.3 Examples of inverse distance models: *IDW* and *GIDS*

The general *IDW* function can be written as (*Jarvis et al., 2001b)*:

$$F(\underline{r}) = \sum_{j=1}^{m} w_i \cdot z(\underline{r}_i) = \frac{\sum_{i=1}^{m} \frac{z(\underline{r}_i)}{|\underline{r} - \underline{r}_i|^p}}{\sum_{j=1}^{m} \frac{1}{|\underline{r} - \underline{r}_j|^p}} \qquad \textbf{(16)}$$

Where w_i are the weights and the various $\underline{r}$ stand for reciprocal distances between the interpolated point and the closest points, m is the number of closest points, p is the power parameter that represents the decay in similarity between values over distance. If $p = 2$, for example, *IDW* turns to be *ISDW*.

The *GIDS* model is nothing but an *MLR* model plus an *IDS* interpolation of the residuals. For a given unmeasured location k and a climatic variable Z, an ordinary least squares regression is performed using the n closest neighbouring locations to calculate coefficients C_x, C_y and C_e representing x, y and *elevation* gradients

$$Z = a + (C_x \cdot X) + (C_y \cdot Y) + (C_e \cdot E) + \varepsilon \qquad \textbf{(17)}$$

Where a is the intercept, ε the residual, X the x-coordinate and Y the y-coordinate for the specified location.

Thus, the basic GIDS formula is:

$$Z_k = \frac{\sum_{i=1}^{n}\left(\frac{Z_1 + \left(C_x \cdot (X_k - X_i)\right) + \left(C_y \cdot (Y_k - Y_i)\right) + \left(C_e \cdot (E_k - E_i)\right)}{d_i^2}\right)}{\sum_{i=1}^{n}\frac{1}{d_i^2}} \qquad \textbf{(18)}$$

Where Z_k is the predicted value at unmeasured location k, Z_i is the measured value at location i and d_i is the distance from measured location i to Z *(Koch et al., 2010; Nalder et al., 1998)*.

3.4.4 Spatial interpolation using *IDW* models in climate studies

The basic *IDW* model was the most used interpolation model until approximately twenty years ago. Afterwards, it was substituted by *ISDW* and recently by *GIDS*.

In literature, *IDW* was applied, e.g., for temperature for Israel by *Kurtzman et al. (1999)*, for maximum temperature and precipitation for Jalisco (Mexico) by *Hartkamp et al. (1999)*, and for a gridded daily dataset of temperature and precipitation for Czech Republic by *Stepanek et al. (2009)*. *ISDW* was applied for rainfall in California by *Curtis et al (1999)*, for the climate model of Iran by *Alijani et al. (2008)*. *ADW* was applied for daily precipitation for Europe by *Hofstra et al. (2008)* and for multivariable maps for the Globe by *New et al. (1999, 2000)*.

GIDS was tested for the first time by *Nalder et al. (1998)* in a study on temperature and precipitation for the Boreal forest in Canada, then it was applied by *Price et al. (2000)* for monthly maps for climate variables for Canada.

IDW was used after *MLR* for residual interpolation for temperature and precipitation in Arizona by *Brown et al. (2002)*. *ISDW* was used after *MLR* for the same reason for temperature for Spain by *Ninyerola et al. (2007)*.

3.4.5 Comparisons between Inverse distance models and similar methods

Basic *IDW* or *IDS* usually produce climatologies with higher statistical errors (e.g. *Ashraf et al., 1997; UnK* vs. *IDW, Zimmerman et al., 1999)*, or similar errors (*Curtis et al., 1999; Jarvis et al., 2001b)* than kriging, but in some cases *IDW* outperforms kriging (*IDW* vs. *CK*,

Hartkamp et al., 1999). Simple *IDW* is usually inferior to *MLR* models with rare exceptions where they lead to similar results (*Goodale et al., 1998)*, and they are rarely compared to splines *(Hartkamp et al., 1999; Jarvis et al., 2001b; Kurtzmann et al., 1999)*, which are usually considered a more accurate technique. Smart distance interpolation is slightly better than *IDS (Pan et al., 2004)*.

GIDS or similar *MLR* plus residual *IDS* interpolation models are considered the best models to realise climatologies. *GIDS* outperforms *ResK, NN, CK, IDS, UnK* and *OK* for temperature and precipitation for *Nalder et al. (1998)*. In particular, authors point out that it is better than kriging methods because it is applied more quickly, easily, and it avoids the subjectivity in the choice of the variogram type from the experimental variogram. For *Kurtzmann et al (1999) MLR* plus *residual IDW* is better than splines or trend surfaces, for *Vicente-Serrano et al. (2003)* it is better than trend surfaces, smart distance searching, *CK*, splines and for *Stahl et al. (2006) GIDS* is better than *NN, MLR, IDS*.

On the other hand, *Price et al. (2000)* prefers *ANUSPLIN* (a splines code) to *GIDS*.

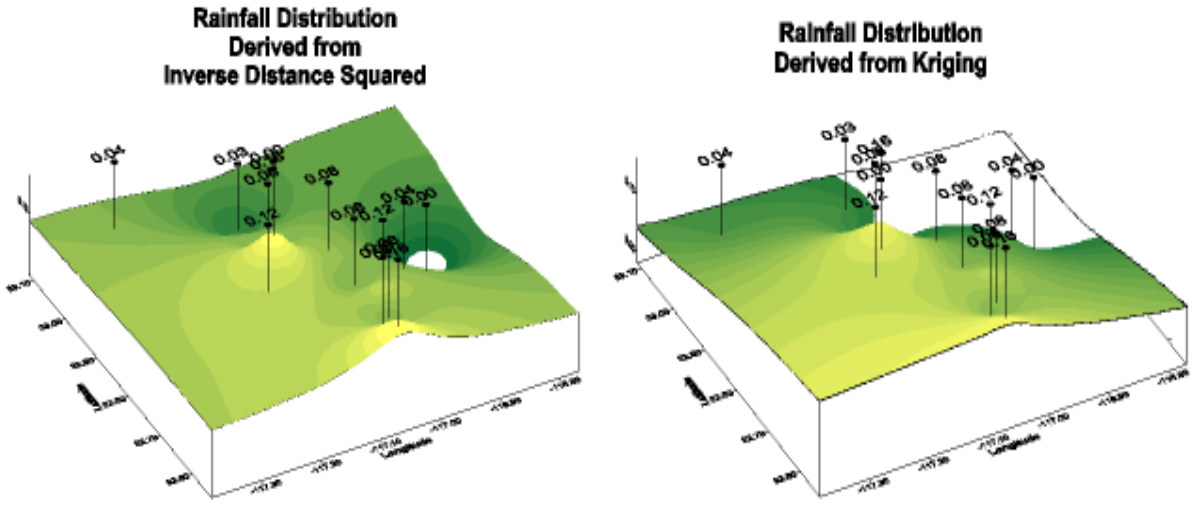

Fig.15 Left: rainfall distribution derived from IDSW; right: rainfall distribution derived from kriging. Data points indicated with black pins. IDSW method can produce bull's eyes but it gives more accurate estimates if data distribution is dense (Curtis et al., 1999)

3.5 Splines models

3.5.1 Definitions and assumptions

In mathematics, a spline is a function defined piecewise by polynomials. In interpolating problems, splines are preferred to polynomials because they yield similar results avoiding the so-called Runge's oscillation problems when high degree functions are used *(Epperson, 1998)*. Then more parameters can be defined in splines, including the amount of "smoothing" *(Hengl, 2009)*, and such smoothing splines prevent over-fitting *(Wiley website, 2010)*.

The smoothing spline function is based on the assumption that there is a measurement error in the data that needs to be smoothed locally *(Hengl, 2009)*; these spline functions for interpolation are normally determined as the minimizers of roughness subject to interpolation constraints. Smoothing splines are considered as a generalization of interpolation splines where the functions are determined to minimize a weighted combination of the average squared approximation error over observed data and the roughness measure.

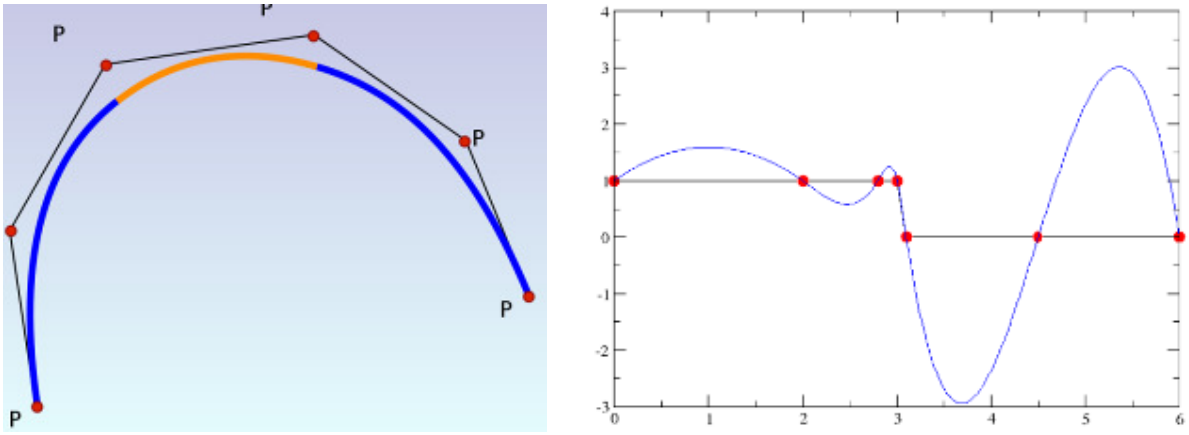

Fig.16 Left: spline of three cubic polynomials; right: mixed linear and cubic spline (Wikipedia, 2010).

3.5.2 Splines models: variants, peculiarities and limits

When an interpolation of densely sampled heights, data or climatic variables is needed, splines perform best. Whilst the main disadvantages are the inability to incorporate a large amount of auxiliary maps to model the deterministic part of the

variation (usually the maximum number is three, in the trivariate splines models), and the fact that smoothing and tension parameters are determined subjectively *(Hengl, 2009)*. Furthermore, in any spline model, the analyst must select the degrees of freedom, the number of knots and decide where they should be placed along the range of x.

Splines can be divided into uniform and not uniform, natural and interpolating and many examples can be found in literature (the name depends on the choice of polynomials type) *(Epperson, 1998; Hengl, 2009; Hutchinson, 1991; Wahba, 1990)*:

- Simple linear regression splines ;
- Quadratic or cubic splines ;
- Natural splines ;
- Bézier splines (Benrstein's polynoms splines) ;
- B-splines ;
- Hermite splines ;
- Cardinal splines ;
- Catmull-Rom splines ;
- Kochanek-Bartels splines ;
- General smoothing and mixed smoothing splines ;
- Thin Plate Splines (TPS) ;
- Multivariate Splines (Bivariate and Trivariate splines) ;
- Regularized splines with tension (and/or smoothing, RSwT) ;

The most used are regularized splines with tension *(Mitasova et al., 1993a)* and *TPS (*e.g. *Hutchinson et al., 2009)*.

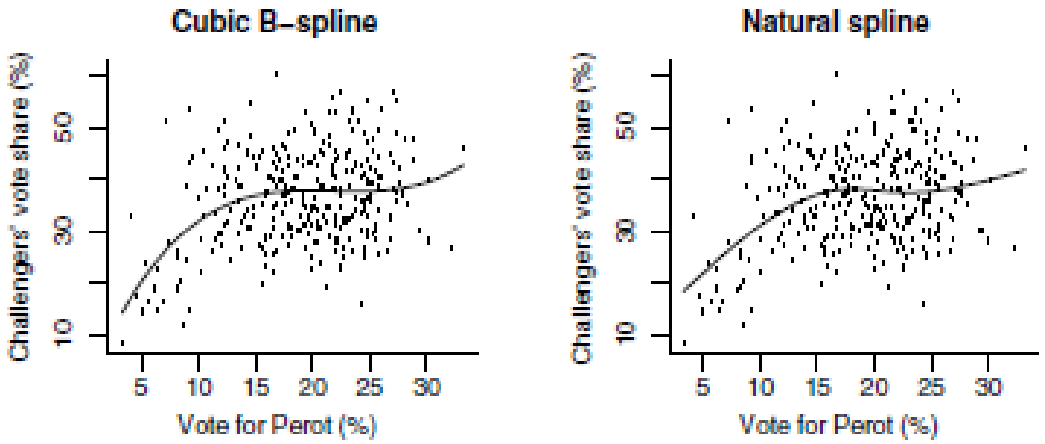

Fig.17 Left: cubic B-spline; right: natural spline. They interpolate a challenger vote's share (from the web)

TPS fits a non-linear flexible n-dimensional surface to data points. The final surface is usually chosen by a routine that virtually removes each data in turn and estimates the deviation of its real value from the value predicted by surface. The final surface is the one that achieves an optimal compromise between minimizing the mean squared deviation summed over all data points and the overall roughness of the estimated surface *(Lennon et al., 1995)*. *TPS* uses a minimal number of co-variables, depending on the smoothing parameter. *TPS* function varies from a surface that exactly interpolates the data to an increasingly smooth function (e.g. a plane). The accuracy of the prediction of the surface (or the trend) strictly depends on the smoothness *(Hancock et al., 2006)*.

Nowadays, the most sophisticated software based on splines is *ANUSPLIN* (*Hutchinson et al. 2009; Milewska et al., 2009)*. It consists in a trivariate (longitude, latitude, elevation) *TPS*. It is a non-parametric generalisation of the multivariate linear regression. It optimises the data smoothing in order to minimize predictive errors by *GCV* (General Cross Validation). Its automatic approach easily adapts to varying station density. It supports calculation of spatially distributed standard errors. It uses a relative scaling of elevation to best represent the elevation impact on surface climate that is more important than horizontal position and it makes use of knots to equi-sample data to fit spatially varying, but regionally defined, elevation lapse rates.

The *ANUSPLIN* software is freely available at *http://fennerschool.anu.edu.au /publications/software/anusplin.php*.

3.5.3 Example of splines model: regularized splines with tension

Following *Hengl (2009)* and *Mitasova et al. (2005)*, in the case of the regularized splines with tension and smoothing, the predictions are obtained by:

$$\hat{z}_S(\underline{s}_0) = a_1 \cdot \sum_{i=1}^{n} w_i \cdot R(v_i) \qquad \textbf{(19)}$$

Where a_1 is a constant and $R(v_i)$ is the radial basis function determined as in equation 20.

$$R(v_i) = -\left[E_1(v_i) + \ln(v_i) + C_E\right] \quad \textbf{(20)}$$

$$v_i = \left[\varphi \cdot \frac{h_0}{2}\right]^2 \quad \textbf{(21)}$$

Where $E_1(v_i)$ is the exponential integral function, C_E is the Euler constant, φ is the generalised tension parameter and $\underline{h}_0$ is the distance between the new and the interpolation point.

The coefficient a_1 and w_i are obtained by solving the system:

$$\sum_{i=1}^{n} w_i = 0 \quad \textbf{(22)}$$

$$a_1 + \sum_{i=1}^{n} w_i \cdot \left[R(v_i) + \delta_{ij} \cdot \frac{\varpi_0}{\varpi_i}\right] = z(\underline{s}_i) \quad ; \quad j = 1,...n \quad \textbf{(23)}$$

Where ω_0/ω_i are positive weighting factors representing a smoothing parameter at each given point $\underline{s}_i$. The tension parameter controls the distance over which the given points influence the resulting surface; the smoothing parameter controls the vertical deviation of the surface from the points.

Regularized spline with tension and smoothing are fairly equivalent to *UnK* and they yield very similar results *(Hengl, 2009)*.

3.5.4 Spatial interpolation using splines models in climate studies

TPS models are often employed in climate interpolation studies. *Hutchinson et al. (1998)* used *TPS* for daily rainfall data of Illinois. *Steinacker et al. (2000)* used *TPS* to interpolate errors from irregularly distributed and noisy datasets. *New et al. (1999, 2000, 2002)* used *TPS* to interpolate global anomalies for many variables. *Hartkamp et al. (1999)* used *TPS* to realize maximum temperature and precipitation climatologies for the Mexican area near Jalisco. *Hancock et al. (2005)* used bivariate splines for temperature for Africa.

The *ANUSPLIN* software was successfully applied to Canadian temperature and precipitation climatologies by *Hutchinson et al. (2009), Price et al. (2000), Milewska et al. (2009)*, to England and Wales for temperature maps by *Jarvis et al. (2001b)* and to global climate surfaces by *Hijmans et al. (2005)* in the version *ANUSPLIN 4.3*.

Splines were also used to interpolate residuals after *MLR*, e.g. by *Lennon et al. (1995)* for temperature for the United Kingdom and by *Ninyerola et al. (2007)* for temperature maps for Spain.

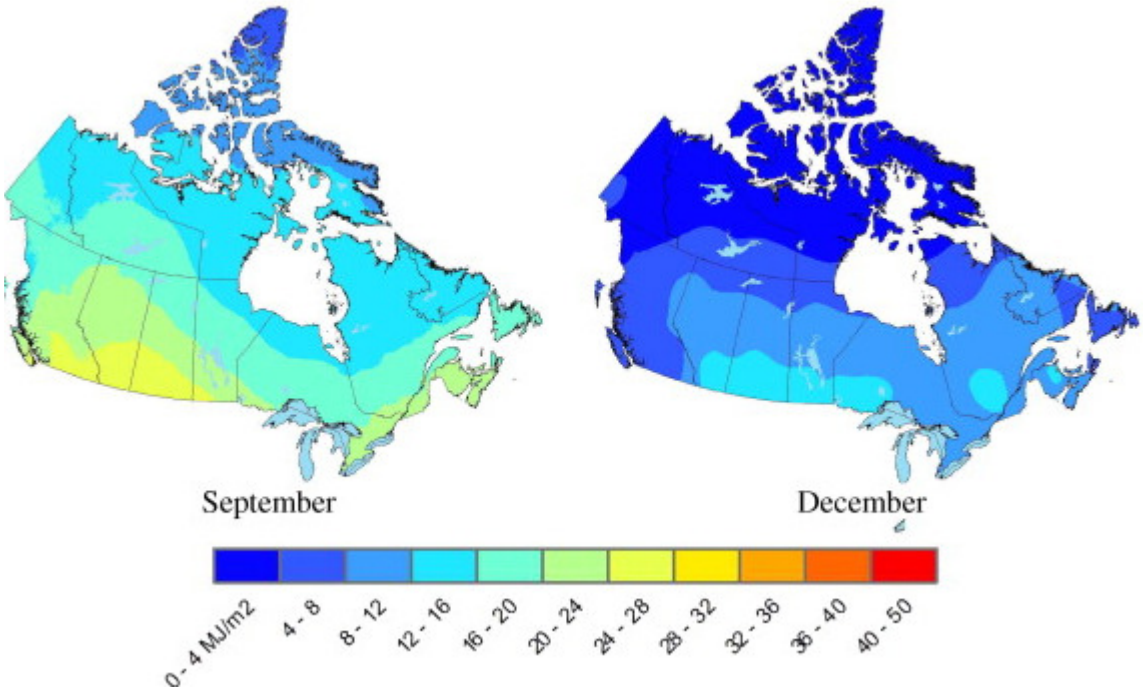

Fig.18 Monthly mean daily insolation (MJ/m^2) for a sun-tracking orientation for September and December in Canada obtained with ANUSPLIN software (McKenney et al., 2008).

3.5.5 Comparisons between splines models and similar methods

In literature, the use of splines models is quite recent. Splines provide better results if compared with trend surfaces, *MLR, IDW* models for *Hijmans et al. (2005)*, with pure *MLR* models for *Jarvis et al. (2001b*), with *IDW* and *CK* for *Hartkamp et al. (1999)*. In combination with *MLR* they are considered the best choice for *Lennon et al. (1995)*. On the other hand, *MLR* and *OK* are preferred to splines by *Chuanyan et al. (2005*), *MLR* are preferred to splines by *Kurtzmann et al. (1999)* and *Vicente-Serrano et al. (2003)* and *GIDS* is preferred to

splines by *Price et al. (2000)*. However, splines are considered equivalent to *IDW* or triangular surfaces by *Robeson (1994)*.

To summarize, the accuracy of the splines accuracy depends on the application.

3.6 *PRISM* model

3.6.1 Definitions, assumptions, and state of evolution

PRISM stands for Parameter-elevation Regressions on Independent Slopes Model (*Daly et al., 1994)*; it is a combined statistical/human-expert approach to climate mapping using *KBS* (Knowledge-Based System) approach (*Daly et al., 1994, Johnson et al., 1997)*.

PRISM uses point data, digital elevation models and other weighted rasters in order to create gridded estimates of daily, monthly or annual climatic variables *(Gibson et al., 1997)*.

The *PRISM* model consists of a local moving-window, climate-elevation regression function that interacts with an encoded knowledge base and inference engine. This knowledge base is a series of rules, decisions and calculations that sets weights for the station data points which enter the regression function. In general, a weighting function contains knowledge about important relationships between the climate field and a geographic or meteorological factor *(Daly et al., 2002)*. It mixes a-priori knowledge of the scientist and a geographical approach. It is *GIS*-compatible and the system of the equations is dynamic, i.e. the model is as open as changeable and developable by users as possible (*Daly et al., 1994; Daly et al., 1998*).

Because of these user-dependent peculiarities, the *PRISM* model needs an objective set of cross validation processes to assess the reliability in every application (*Daly et al., 2002)*.

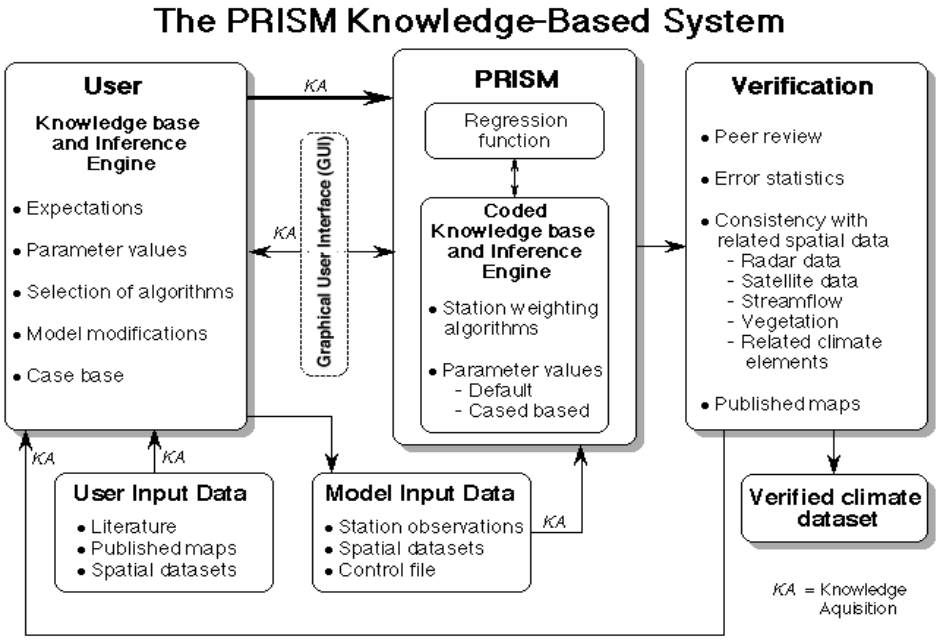

Fig:19 The PRISM KBS conceptual scheme (Daly et al., 2002)

3.6.2 *PRISM*: peculiarities, merits and limits

PRISM was originally created as a spatialization model for precipitations and it is uniquely suited with mountain regions. It incorporates a conceptual framework that allows for spatial scale and pattern of orographic precipitation to be quantified and generalised. However, it was further adapted to temperature, snow fall, growing degree days, weather parameters, facet grids and so on *(Gibson et al., 1997; Daly et al., 1997; Johnson et al., 1997; Taylor et al., 1997)*.

PRISM is able to extrapolate over large elevation ranges to reproduce gradients between different climatic regimes, to understand the terrain induced climate transitions (as topographic slopes, barriers, facets and so on), to include coastal effects, a two-layers atmosphere representation, the orographic effectiveness of the ground and so on *(Daly et al., 1997; Daly et al., 2002)*. It is based on the assumption that, for a localized region, the most important factor in the distribution of most climate elements is elevation. Thus, the governing equation is a weighted linear elevation regression function and it is necessary to place bounds on the regression slope because sparse station data may not accurately represent the local climate-elevation relationship, especially if only a narrow range of elevation is available.

Upon entering the regression function, a weight is assigned to each station. The weight is based on several factors, e.g. an aspect weight, a radial distance weight, a coastal proximity weight and so on. The weights can be based on a-priori hypothesis and they can even be refined with iterative processes that choose the final weight as they minimize the mean absolute error and the root mean squared error. The station weights are a product of sub-weights and the number and kind of weights vary from one *PRISM* application to another. *PRISM* algorithms are in a state of constant development (*Daly et al., 2002).*

The main advantage of this methodology is the chance to continuously change weights during the analysis, adapting them to data or refining them after new analysis. Such weights are based on strong physical motivations because they are based on expertise knowledge of climatologists.

On the other hand, the main limit of this technique is the not-so-easy reproducibility of the *PRISM*. It varies from one region to another and it strongly depends on the climate variable modelled.

A modified version of the *PRISM* used by *Daly et al. (2002)* for the USA is used to realise precipitation climatologies for Italy by University of Milan and *ISAC-CNR* of Bologna.

3.6.3 Example of *PRISM*: the model used by *Daly et al. (2002)* for the USA

The governing equation, that is the elevation regression function, is a simple linear regression:

$$Y = \beta_1 X + \beta_0 \qquad \beta_{1m} < \beta_1 < \beta_{1x} \tag{24}$$

Where Y is the predicted climate element, β_1 and β_0 are the regression slope and intercept, respectively, X is the DEM elevation at the target grid cell, β_{1m} and β_{1x} are the minimum and maximum allowable regression slopes. Such an equation is developed from x, y pairs of elevation and climate observations supplied by station data.

All stations located within a user-specified maximum radius *r* from the target cell are used to get the regression and the *r* value is set to get a compromise between statistical robustness and desire for local detail.

Then, a weight is assigned to each station and the combined weight is given as follows:

$$W = \left[F_d \cdot W(d)^2 + F_z \cdot W(z)^2\right]^{\frac{1}{2}} \cdot W(c) \cdot W(l) \cdot W(f) \cdot W(p) \cdot W(e) \qquad \textbf{(25)}$$

Where *W(d)* is the distance weight, *W(z)* is the elevation weight, *W(c)* is the cluster weight, *W(l)* is the vertical layer weight, *W(f)* is the topographic facet weight, *W(p)* is the coastal proximity weight, *W(e)* is the effective terrain weight, F_d and F_x are the distance and elevation weighting importance factors.

All the weights and importance factors, individually and combined, are normalized to sum to unity.

The distance weight is given as:

$$W(d) = \begin{cases} 1 & d = 0 \\ \dfrac{1}{d^a} & d > 0 \end{cases} \qquad \textbf{(26)}$$

Where *d* is the horizontal distance between the station and the target grid cell, *a* is the distance weighting exponent. If *a = 2* this weight is equivalent to a *IDSW* model.

The elevation weight is given as:

$$W(e) = \begin{cases} \dfrac{1}{\Delta z_m^b} & \Delta z \le \Delta z_m \\ \dfrac{1}{\Delta z^b} & \Delta z_m < \Delta z < \Delta z_x \\ 0 & \Delta z \ge \Delta z_x \end{cases} \qquad \textbf{(27)}$$

Where Δz is the absolute elevation difference between the station and the target grid cell, b is the elevation weighting exponent, Δz_m is the minimum elevation difference, Δz_x is the maximum elevation difference.

The cluster weight seeks to down-weight stations located in tight clusters, in order to minimize over-representation of one location over others in the regression function.

The facet weight (also called aspect weight) is given as:

$$W(f) = \begin{cases} 1 & \Delta f \le 1 \quad and \quad B = 0 \\ \dfrac{1}{(\Delta f + B)^c} & \Delta f > 1 \quad or \quad B > 0 \end{cases} \qquad \textbf{(28)}$$

Where Δf is the absolute orientation difference between the station and the target grid cell, B is the number of intervening barrier cells with an orientation significantly different than that of the target grid cell, c is the facet weighting exponent.

The coastal proximity weight is given as:

$$W(e) = \begin{cases} 1 & \Delta p = 0 \\ 0 & \Delta p > p_x \\ \dfrac{1}{\Delta p^v} & 0 < \Delta p \le p_x \end{cases} \qquad \textbf{(29)}$$

Where Δp is the absolute difference between the station and the target grid cell coastal proximity index, v is the coastal proximity weighting exponent, p_x is the maximum proximity difference. Usually $v = 1$ for regions where coastal effects are significant.

The vertical layer weight is given as

$$W(l) = \begin{cases} 1 & \Delta l = 0 \quad or \quad \Delta z \le \Delta z_m \\ \dfrac{1}{(\Delta z - \Delta z_m)^y} & \Delta l = 1 \quad and \quad \Delta z > \Delta z_m \end{cases} \qquad \textbf{(30)}$$

Where Δl is the absolute layer difference between the station and the target grid cell (1 for adjacent layer, 0 for same layer), y is the vertical layer weighting exponent.

For the effective terrain weight see *Daly (2002)*.

3.6.4 Spatial interpolation using *PRISM* models in climate studies

PRISM was conceived by Christopher Daly, together with his research group, in the '90s. Then it was established the "*PRISM* Group" at the Department of Geosciences and College of Oceanographic and Atmospheric Sciences of the Oregon State University (in Corvallis). The complete list of reports and papers on *PRISM* of the *PRISM* Group can be found at *http://www.prism.oregonstate.edu/docs/index.phtml*.

The *PRISM* group applied *PRISM* to minimum and maximum temperature and precipitation climatologies for Oregon, for the conterminous United States, for China, for Alaska and Canada, Puerto Rico (*Daly et al., 1994; Daly et al., 2001; Daly et al., 2003; Daly et al., 2009; Johnson et al., 1997; Kittel et al., 1997; Parzybok et al., 1997)*. They used *PRISM* to quality-check data and to create climatic databases that were further used by other research groups, e.g. in *Wang et al. (2000)* and in *Hamann et al. (2005)*.

The most accurate version of *PRISM* was used to create physiographically-sensitive maps, for 1971-2000, for temperature and precipitation across the conterminous USA (*Daly et al., 2009)*. It uses a hi-res *DEM*, a topographic facets grid, a moist boundary layer height grid, a topographic index grid, an effective terrain height grid, a US boundary high-resolution grid, a recalculated coastal proximity grid. The model is based on a linear elevation regression plus a cluster weight, a distance weight, an elevation weight, a topographic facet weight, a coastal proximity weight, a two-layer atmosphere weight, a topographic position weight, an effective terrain weight (all of these weights were improved from weights used in *Daly et al. (2002)*, see Chapter 3.6.3).

In Europe, modified and simplified versions of *PRISM* were used to create precipitation climatologies for Switzerland, as in the Hydrologic Swiss Atlas *(Frei et al., 1998; Schwarb et al., 2000; Schwarb et al., 2001)*, and to realise northern Italian precipitation climatologies by (*Brunetti et al., 2009c; Brunetti et al., 2010*). Such Italian climatologies were completed receiving help in the framework of this PhD project.

Because of its particular formalism and its recent creation, *PRISM* was applied in a small number of climate spatialization studies, always providing satisfactory results in

terms of statistical parameters and in cross validation procedures. Nevertheless, a few comparisons with similar models were realised (e.g., *ANUSPLIN* vs. *PRISM* for Canada and Alaska, *Simpson et al., 2005*).

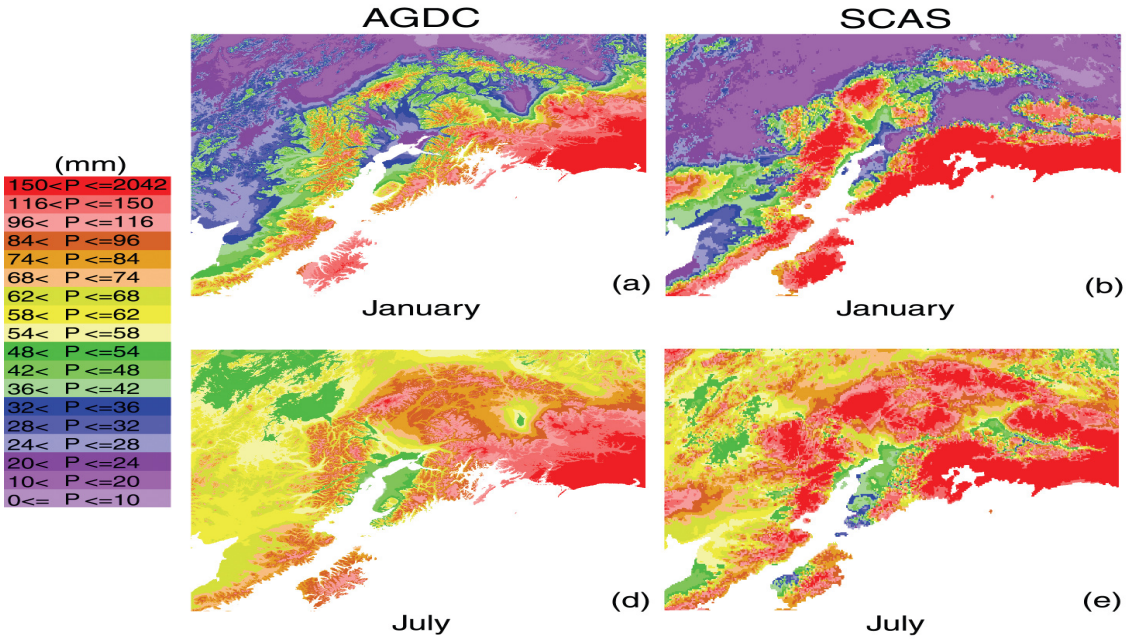

Fig.20 Precipitation January and July maps for Alaska obtained with ANUSPLIN (left) and by PRISM (right). PRISM maps are based on a denser dataset, the Oregon State University's Spatial Climate Analysis Service (SCAS) , splines maps are based on the Alaska Geospatial Data Clearinghouse (AGDC) (Simpson et al., 2005).

3.7 Other spatialization models

Besides kriging models, regression models, inverse distance models, splines models and *PRISM*, other spatialization models can be found in literature.

3.7.1 Nearest neighbours models

Nearest neighbours (*NN*) (also known as point sampling or Thiessen polygons) is the simplest interpolation method; it assigns to the target point in space the value of the nearest sampled point, without considering any other sampled point *(Bremner et al., 2005)*. The only advantage is the very simple algorithm and the fast encoding of the model.

The most common variants are k-nearest neighbour algorithm (*KNN*), approximate *NN*, *NN* distance ratio, and "all *NN*" method.

In climate studies, this technique is rarely used because it provides inaccurate results. In recent literature, it is only used in methods comparisons *(Nalder et al., 1998)*.

NN is connected to Voronoi diagrams. A Voronoi diagram is a decomposition of space into cells, one for each given point, this is equivalent to *NN* because the Voronoi diagram assigns to every point of the cell the value of the sampled point. This method is also known as Dirichlet tessellation.

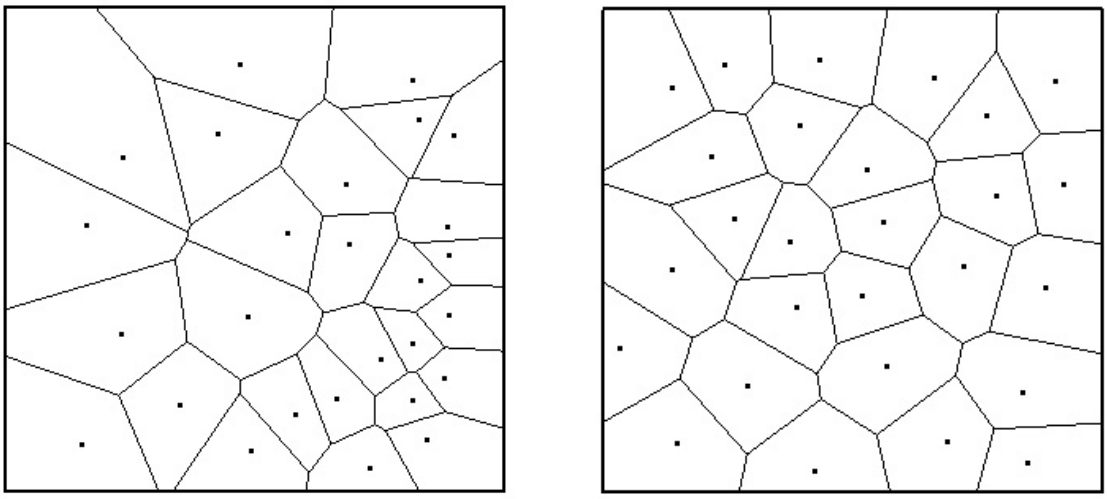

Fig.21 Examples of Voronoi diagrams (left and right).

3.7.2 Trend surface models

Assuming that the values of a target variable at some location are a function of coordinates, we can determine its values by finding a function which passes through or close to the given set of discrete points *(Hengl, 2009)*. If the function is fitted for the whole dataset (trend), this method is called trend surfaces (*TS*), whilst it is called moving surface interpolation if the function is fitted for a local moving neighbourhood *(Hardy, 1971)*.

Trend analysis deals with the measurement of trends, which can be represented by lines, surfaces of hyper-surfaces. The technique is designed to separate an observed contour map surface into regional and residual components. The regional component is obtained by fitting a low order surface to the data using, e.g., least squares techniques,

thus the regional part can be reasonably approximated by a smooth continuous surface that is usually expressed by polynomials *(Wren, 1973)*. If higher order polynomials are introduced, many artefacts and a serious local overshooting of the values can be caused *(Hengl, 2009)*.

The main limits of *TS* models are: the choice of the low order surface equations is made by the user, a moving surface often fails to represent discrete changes and *TS* does not rely on empirical analyses of data *(Diggle et al., 2007)*.

Trend surfaces are rarely employed in climate modelling, but they usually produce less accurate results than splines, kriging, *IDW* and regression models *(Stahl et al., 2006; Kurtzman et al., 1999; Jarvis et al., 2001b)*.

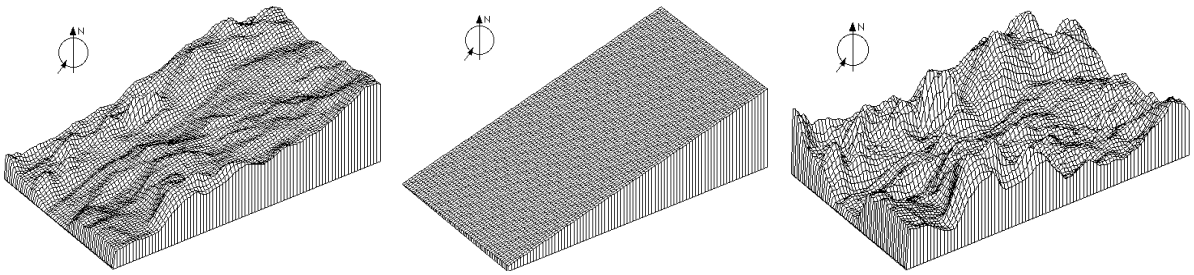

Fig.22 A trend surface decomposition: left, the physical surface; centre, the first order trend; right, the first order residual. Images are referred to Bouguer gravity, N-W Kansas. (Kansas Geological Survey website)

3.7.3 Optimal interpolation models

Optimal interpolation (*OI*) is an old technique for four-dimensional data assimilation; it is a statistical approach developed by Lev Gandin *(Gandin, 1963)* and afterwards applied especially in meteorology.

Qualitatively, the Optimal Interpolation works as follows: at each grid point where *OI* wants to estimate a value, a number of surrounding data points is sampled. Based on the distance between these points, their distances from the grid point, and their error values, each sampled data point is assigned an "alpha" weight. The background field at each data point is then subtracted from each data value. These new values are multiplied by their alpha weights and then added together (in effect providing a "weighted average" of the deviation of each data point from its associated background value). This value plus

the background field at the grid point in question is the final value calculated by the *OI* routine. Important parameters in *OI* are the number of points used in calculations, radius of influence, correlation length, and error *(Polar Science Center, Washington, website)*.

OI is a simplified version of ordinary kriging and it was most used in the 1980s and 1990s. It then fell from grace and was substituted by kriging variants, though in some recent papers it is still used and some meteorological institutes still use it *(Sen et al., 2001a; NOAA website; Hartkamp et al., 1997; Reynolds et al., 1994; Pasini et al., 2006)*.

3.7.4 Neural networks models

In climate variables interpolation, when we refer to neural networks, we mean artificial neural networks (*ANN*), i.e. networks composed of artificial neurons or nodes.

An *ANN* is a mathematical model that simulates the structure or the functionalities of biological neural networks. This method consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation.

Usually, *ANN* is an adaptive system that changes its structure based on external or internal information, which flows through the network during the learning phase. Modern neural networks are non-linear statistical data modelling tools and they are used to model complex relationships between inputs and outputs or to find patterns in data *(Gurney, 1997; Smith, 1993)*.

Variants of *ANN* are, for example, feed-forward neural network, radial basis function (*RBF*) network, Kohonen self-organizing network, recurrent network, simple recurrent network, Hopfield network, echo state network, long short term memory network**,** stochastic neural networks, Boltzmann machine**,** modular neural networks, committee of machines, associative neural network, physical neural network, instantaneously trained networks, spiking neural networks, dynamic neural networks, cascading neural networks, neuro-fuzzy networks, compositional pattern-producing networks an so on *(Haykin, 1999; Bishop, 1995)*.

The main disadvantage of this technique is the very complicated software needed to perform large dataset interpolations and predictions.

One of the most recent applications of neural networks is in climate research where *ANN* are used in function approximation or in regression analysis and even in time series prediction or in climate change scenarios *(Gardner et al., 1998; Hsieh et al., 1998; Knutti et al., 2003; Antonic et al., 2001).*

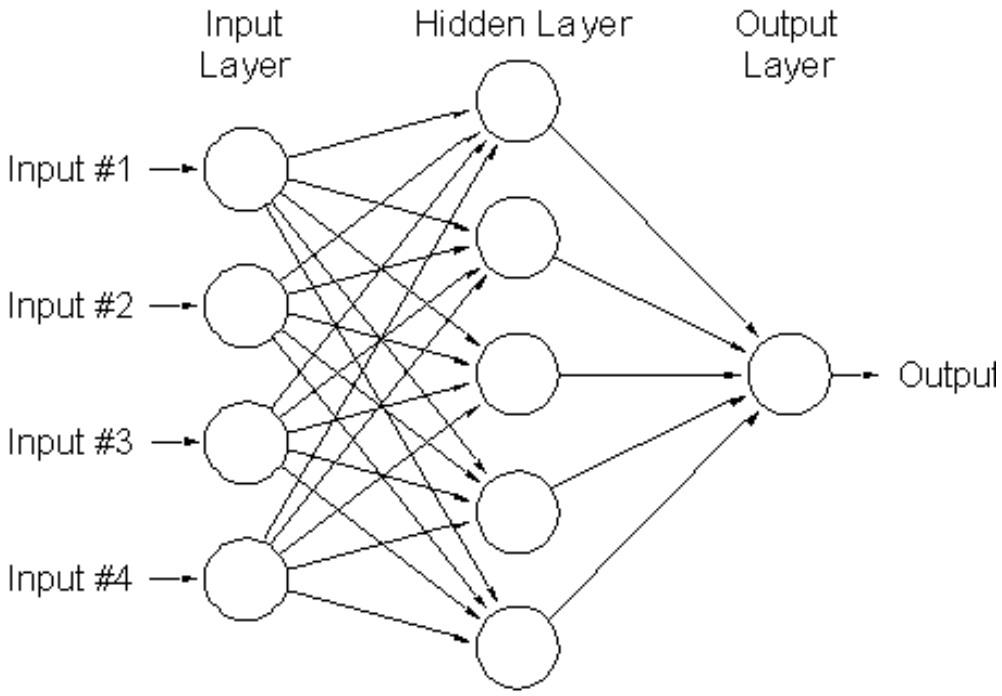

Fig:23 Common scheme of a neural network: from input layer towards hidden layer to output layer.

3.8 Validation of a spatial interpolation model

Spatial interpolation models produce results that must be validated. Statistical tests evaluate the robustness of the output by means of re-sampling techniques and/or statistical parameters, thus single and overall errors are quantified and models can be considered more or less accurate, depending on the purpose and on the established error thresholds.

The vast majority of the climate models base their validation on the evaluation of some difference parameters. *ME* (Mean Error, or *SE*, Standard Error), *RMSE* (Root Mean Squared Error), *MAE* (Mean Absolute Error) and related quantities are calculated using

the differences between the measured data and the modelled ones. Each predicted value is compared with the corresponding station value. The final cited parameters are obtained as the overall average of single data error parameters.

Recently, parametric and non-parametric statistical tests were applied in order to validate climate models. The most common are cross validation, jack-knife and bootstrap. They usually consist of dividing the input dataset in subsets (cross validation) or in leaving one data out of the analysis (jack-knife), recalculating statistical parameters without using a subset or a single data, and finally comparing measured values and predicted values of the data excluded during the realization of the model. Other less-used techniques are permutation tests, exact tests and similar tests.

A model can also be validated by comparing outputs (in grids or maps) with other published models which have already been validated. However, this process is difficult because climate models are not so easily compared because of data formats, different digital elevation models, different grid cell sizes, unavailability of models for some areas and so on.

Nevertheless, even if statistical parameters and re-sampling techniques are an objective way of proving the validity of a climate model, local researchers' knowledge and experience should also be used to get a stronger validation *(Daly et al., 2002)*. Especially when datasets are not very dense, problems like local unreasonable anomalies can be better solved using a-priori knowledge other than statistical tests.

3.8.1 Statistical parameters for prediction errors

In literature, a consistent part of climatologies are validated only by means of statistical parameters as *r*, r^2, *ME*, *MAE*, *RMSE*, without re-sampling techniques (e.g., *Vicente Serrano et al., 2003; Goodale et al., 1998; Brown et al., 2002; Stahl et al., 2006; Benavides et al., 2007; Atkinson et al., 2002; Alsamamra et al., 2009)*. These parameters are calculated by averaging all the single parameters for each input data versus predicted data.

Co-variation or correlation between an observed data and a model-predicted data are usually evaluated by means of Pearson's product-moment correlation coefficient (*r*) and of

the coefficient of determination (r^2), in which is embodied the proportion of the "variance explained" *(Fox, 1981; Willmott, 1982)*, especially in regression models.

It was shown that r and r^2 are insufficient in making meaningful distinctions between models (*Powell, 1980*). Statistically significant values of such parameters may be misleading because they are often unrelated to the sizes of the differences between O and P *(Willmott et al., 1985)*, where the observed value is defined as O and the predicted value is defined as P. Thus, r and r^2 are parameters statistically insignificant and they have no real practical value in the evaluation of a model performance *(Willmott, 1982; Wilmott et al., 1985)*.

Six types of "difference measures" are worth being listed. They are all based on the difference between observed data and predicted data.

Let us also define the number of observations as N, the mean of observed values as $\overline{O}$, the mean of model-predicted values as $\overline{P}$, $P_i - \overline{O} = P_i^{'}$ and $O_i - \overline{O} = O_i^{'}$ *(Vicente-Serrano et al., 2003)*.

Mean Bias Error (*MBE*) (alternatively called Mean Error or *ME*) quantify bias *(Fox, 1981)* and can be written as:

$$MBE = \frac{1}{N} \cdot \sum_{i=1}^{N} (P_i - O_i) \qquad \textbf{(31)}$$

Mean Absolute Error (*MAE*) (alternatively called Standard Error, *SE (Fox, 1981)*) describes the average difference and it can be written as:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} \left| (P_i - O_i) \right| \qquad \textbf{(32)}$$

Root Mean Squared Error (*RMSE*) quantifies the average difference in a different way *(Fox, 1981)* and can it be written as:

$$RMSE = \left(\frac{1}{N} \cdot \sum_{i=1}^{N} (P_i - O_i)^2 \right)^{\frac{1}{2}} \qquad \textbf{(33)}$$

RMSE and *MAE* are among the best overall measures of model performance, but *MAE* is less sensitive to extreme values and it avoids the physically artificial exponentiation. Such parameters are similar and they should be reported together.

The **variance of the distribution of differences** (s_d^2) can be seen as the unbiased difference or the average "noise" *(Fox, 1981)* and it can be written as:

$$s_d^2 = MAE - \left(\overline{P} - \overline{O}\right)^2 \tag{34}$$

The **index of agreement** (*d*) is a descriptive parameter, useful to cross compare models *(Willmott, 1982)* and it can be written as:

$$d = 1 - \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left(\left| P_i' \right| + \left| O_i' \right| \right)^2} \right] \quad ; \quad 0 \le d \le 1 \tag{35}$$

The **Model Efficiency** (*EF*) *(Greenwood et al., 1985)* is an indicator of model intrinsic limitations and it can be written as:

$$EF = 1 - \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} \left(\overline{O} - O_i \right)^2} \right] \tag{36}$$

If *EF* is close to zero, the mean value of the observations is more reliable than the predictions and it suggests that model should be improved.

Other two indices ought to be listed *(Willmott, 1982)*, i.e. the **systematic mean squared error** (MSE_s) can be written as:

$$MSE_s = \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \left(\hat{P}_i - O_i \right)^2 \right) \tag{37}$$

Where $\hat{P}$ is defined as the average of the predicted values.

The **unsystematic mean squared error** (MSE_u) can be written as:

$$MSE_u = \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \left(P_i - \hat{P}_i \right)^2 \right) \quad \textbf{(38)}$$

The system is conservative, consequently:

$$MSE_s + MSE_u = MSE \quad \textbf{(39)}$$

Where MSE is the **Mean Squared Error**.

Difference parameters for multivariable cases can be found in *(Willmott et al., 1985)*.

Willmott (1982) suggests that the interpretation of statistical parameters should be descriptive, based on scientific grounds and physically reasonable, they should not rely only on the basis of the statistical significance of the measurements. Furthermore, summary statistics and graphics can also help in understanding the accuracy of a model.

3.8.2 Bootstrap

The bootstrapping (or bootstrap) is a re-sampling technique that estimates the precision of sample statistics as medians, variances, and percentiles by drawing randomly with replacement from a set of data points.

The properties of an estimator (e.g. its variance) are estimated by measuring those properties when sampling from a "population" distribution. If the dataset is made of independent and identically distributed data, this can be implemented by creating a number of re-samples of the dataset; each of them is obtained by random sampling with replacement from the original dataset.

The bootstrapping should be used when the theoretical distribution of a statistic is unknown, when the sample size is insufficient for straightforward statistical inference or when power calculations have to be performed and only a small pilot sample is available *(Adèr et al., 2008)*. The bootstrapping can also be used for constructing hypothesis tests.

The main advantage consists in its simplicity to derive error estimates, the main disadvantage is that it does not provide general finite-sample results *(Efron, 1979)*.

The bootstrap is similar to jack-knife and to cross validation, but jack-knife is commonly used to estimate biases and variances of sample statistics and cross validation applies to a subsample the parameters calculated from another subsample.

There are many bootstrapping schemes: case re-sampling, smooth, parametric, wild and so on *(Davison et al., 1997)*.

Let us consider a dataset made of a random sample of size n from an unknown probability distribution F on the real line *(Efron et al., 1983)*:

$$X_1, X_2, ..., X_n \approx F \tag{40}$$

Having observed $X_1 = x_1$, $X_2 = x_2$, …, $X_n = x_n$, we compute the sample average:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{41}$$

And we use it as an estimate of the average of F.

The dataset provides also the *RMSE* of estimated average, i.e. the standard error of $\overline{X} = \bar{x}$:

$$\hat{\sigma} = \left[\frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right]^{\frac{1}{2}} \tag{42}$$

The bootstrap generalises equation (42) within the next steps: let $\hat{F}$ be the empirical probability distribution of the data, we extract a random sample from $\hat{F}$:

$$X^*_1, X^*_2, ..., X^*_n \approx \hat{F} \tag{43}$$

Each X^*_i is drawn independently with a replacement and with equal probability from the set $x_1, x_2, \ldots, x_n$, then we have an average:

$$\overline{X}^* = \sum_{i=1}^{n} \frac{X^*_i}{n} \tag{44}$$

And a variance of:

$$var\,\overline{X}^{*} = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textbf{(45)}$$

The **boostrap** estimate of standard error for an estimator $\hat{\theta}(X_1, X_2, ..., X_n)$ is:

$$\hat{\sigma}_{BOOTSTRAP} = \left[\mathrm{var}\,\hat{\theta}(X_1^*, X_2^*, ..., X_n^*\right]^{\frac{1}{2}} \qquad \textbf{(46)}$$

It is evident that:

$$\left[\frac{n}{n-1}\right]^{\frac{1}{2}} \cdot \hat{\sigma}_{BOOTSTRAP} = \hat{\sigma} \quad for \quad \hat{\theta} = \overline{X}\,. \qquad \textbf{(47)}$$

In recent published climate maps, the bootstrapping validation is rarely used. An example can be found in *Mullan et al. (2006)* for New Zealand's temperature and precipitation anomalies.

3.8.3 Jack-knife

The jack-knifing process is similar to the bootstrapping. The basic idea of the jack-knife is the systematically re-computing of the statistic estimates (e.g. *MAE, ME, RMSE*) by leaving out one data at a time from the input dataset, thus the statistic parameters are calculated for the new *n* datasets, then globally averaged.

Jack-knife is a less general technique than bootstrap and it is applied more easily to complex sampling schemes, but the computation of jack-knife requires more time. If we apply the bootstrap repeatedly on the same dataset, it yields slightly different results from time to time, as opposed to the jack-knife, which yields every time the same results if the subsets to be removed are the same *(Efron, 1981)*.

Let us go back to the example given in Chapter 3.8.1 and let us recalculate equation (42) using a jack-knife point of view *(Efron et al., 1983)*.

We can write the sample average of the dataset, deleting the *nth* point as:

$$\bar{x}_{(i)} = \frac{1}{n-1} \cdot \sum_{j \neq i} x_i \tag{48}$$

And the average of the deleted averages as:

$$\bar{x}_{(\cdot)} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{49}$$

Thus, the **jack-knife** estimate of standard error is:

$$\hat{\sigma}_{JACKKNIFE} = \left[\frac{n-1}{n} \cdot \sum_{i=1}^{n} \left(\bar{x}_{(i)} - \bar{x}_{(\cdot)} \right)^2 \right]^{\frac{1}{2}} \tag{50}$$

It can be verified that equation (50) and equation (42) are identical, but equation (50) can be applied to any estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$, provided that we replace $\bar{x}_{(i)}$ with $\hat{\theta}_{(i)} = \hat{\theta}(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$ and $\bar{x}_{(\cdot)}$ with $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{(i)}}{n}$.

Jack-knife is used more than bootstrap but far less than cross validation in order to validate climate models. An example is provided by *Jarvis et al (2001b)* for the comparison between different interpolation methods for temperature climatologies of the United Kingdom.

3.8.4 Cross Validation

Cross validation is also called rotation estimation and is a technique used for assessing how the results of a statistical analysis can be generalised to an independent dataset and it is used to estimate how accurately a predictive model will perform in practice.

Cross validation consists in partitioning a sample of data into complementary subsets, performing the analysis on one subset (the training set) and then validating the analysis on the other subset (the testing set). Many rounds of cross validation, with different partitions, are usually performed and then the statistical parameters calculated from single rounds are globally averaged *(Kohavi, 1995)*. Cross validation should be used only if the validation set and the test set are drawn from the same population. Cross validation is also used to select and remove predictors in climate models because it is a valuable tool against over-fitting and over-sampling problems *(Picard et al., 1984)*.

There are many different variants of cross validation: repeated random sampling, k-fold (the original dataset is partitioned in k subsets), double k-fold, leave one out (that is jack-knife).

Let us go back once again to the example given in Chapter 3.8.1 and let us recalculate equation (42) using a cross validation point of view *(Efron et al., 1983)*.

Suppose we decide to predict a new observation from F, see equation (41), and we call it X_0, using the estimator $\overline{X}$ as a predictor.

The expected squared error of the prediction is

$$E\left[X_0 - \overline{X}\right]^2 = \left(\frac{n+1}{n}\right) \cdot \mu_2 \qquad \textbf{(51)}$$

Where μ_2 is the variance of the distribution *F*.

Thus, an unbiased estimate of $\left(\frac{n+1}{n}\right) \cdot \mu_2$ is:

$$(n+1) \cdot \hat{\sigma} \qquad \textbf{(52)}$$

Cross validation is a way to obtain nearly unbiased estimators of prediction errors in more complicated situations. First, we delete x_i data from the dataset, one a time. Second, we recalculate the prediction rule on the basis of the *n-1* points. Third, we evaluate if the recalculated rule satisfactorily predicts the deleted point by means of statistical parameters (such as *ME, MAE, RMSE* and so on). Fourth, we average predictions over all deletions of and x_i. Usually, to form a subset, many data are taken apart from the calculations.

In the simple case showed above, the **cross-validation** estimate of prediction error is:

$$\hat{\sigma}_{CROSSVALIDATION} = \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}_{(i)} \right)^2 \right]^{\frac{1}{2}} \quad \textbf{(53)}$$

Cross validation and its variants are probably the most used techniques in validation of temperature, precipitation and other climate interpolated variables *(Lennon et al., 1995; Daly et al., 2003; Gyalistras, 2003; Agnew et al., 2000; Holdaway, 1996; Hancock, 2005; Kurtzman et al., 1999; Daly et al., 2009)*.