Spatial statistics overview

Spatial interpolation

Spatial interpolation can be described as the process of estimating the value of a variable at an unobserved location based on the values of the variable at surrounding observed locations. These locations are typically represented as equal distance square or hexagon shaped grids, but other representations like polygon centroids and points could also be used. These representations typically have higher data density requiring more computational time for the predictions. Spatial interpolation methods can be classified as global or local which separately can be considered deterministic or stochastic and produce either exact or approximate results.

Global and local

The phrases local and global define how geographically distributed observations are used by the interpolation method. In global interpolation a single function is applied to all receptors in a selected region. In contrast, local interpolation uses samples of known observations to estimate a section of receptors where the surface is built from these different sections.

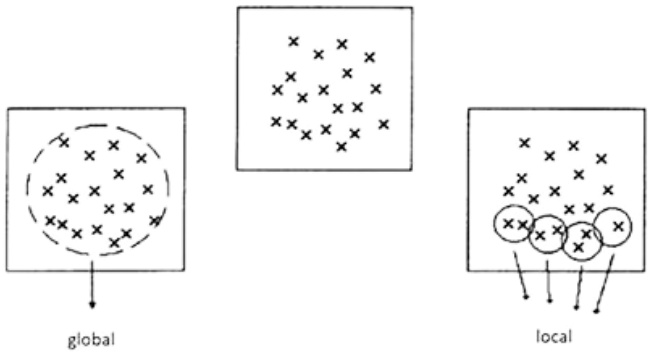

Fig 1.1.1: Sampling of observation using global and local interpolation methods.

Global interpolators aim to evaluate the global spatial structure and in general, provide a surface with few sudden changes since they use the averaging principle. Global interpolation usually results in smoother surfaces which lessens the impact of extreme observations. When the surface being modelled is known to have an overall trend, such methods are better suited.

Local interpolators are more suited when there is little or no knowledge about the overall trend of the surface you are attempting to model and aim to capture local or short-range variations. For many meteorological variables, using local methods to estimate unknown values at a point is more accurate than using global methods. (Li Y. et al 2016).

Deterministic and stochastic

Deterministic interpolation methods are based on precisely specified spatial relationships, such as similarity and smoothness. These interpolation methods are suited if you know enough about the surface you are modeling, to describe its properties as a mathematical function. Unfortunately, this is rarely the case for surfaces with real-world features.

Stochastic methods consider the concept of randomness. The interpolated surface could be one of many observable surfaces, all of which could have produced known data points. Stochastic interpolation methods include trend surface analysis which can calculate the statistical significance of the surface and the uncertainty of the receptor variables.

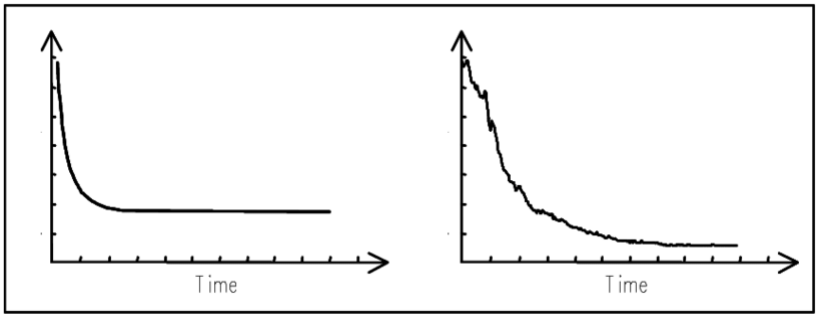

Fig 1.1.2: Comparison of deterministic function (left) and stochastic function (right). The stochastic function has some degree of randomness in its predictions.

Deterministic interpolation methods do not allow for error estimation of receptor variables. Stochastic methods deal with the randomness of receptors and the associated errors. In these methods, uncertainty is displayed as estimated variance (Scheuerer M. et al 2013).

In some instance both deterministic and stochastic methods could be used in so-called combined models. (Li J. et al 2014).

Exact and approximate

Exact interpolation methods consider all observations which are available. This means that the generated surface must pass exactly through all the observations and not smooth or alter their values. These interpolation methods are best suited when the observed variables give a high level of confidence in the measurements made (ex: in a depth map it does not make sense to generate a surface where observations have a different variable than was measured).

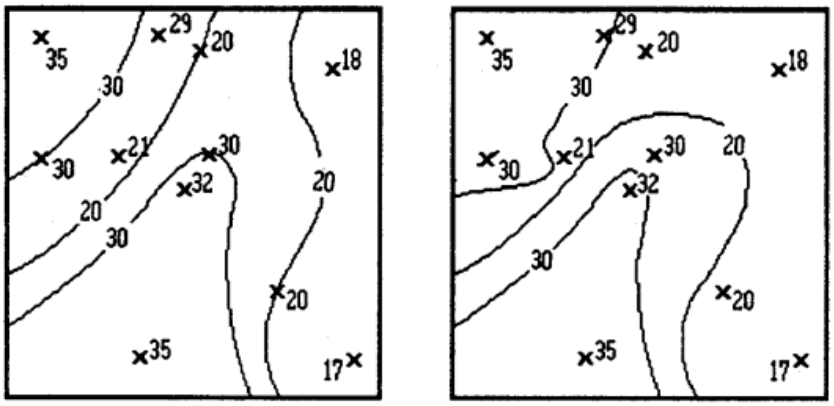

Fig 1.1.3: Exact (left) and approximate (right) interpolators. The left figure shows that the generated surface passes through all observations whilst the right figure shows some observation fall out of the generated surface.

Approximate interpolation methods do not have to consider the observations but can smooth or alter the data points to fit general trends. Approximate methods are more appropriate when there is some uncertainty in the measurements made at the sampled points or when deviations are permitted. It can be assumed that many datasets have slowly changing global trends and that rapidly changing local fluctuations add up to introduce uncertainty (error) into the recorded values. The effect of smoothing is therefore to reduce the effect of errors on the resulting surface.

Geographic information system

A geographic information system (GIS) is a computer system that captures, stores, verifies, displays, and defines how spatial data relates to locations on the Earth's surface using a coordinate reference system. It can be used to analyze relationships between different spatial features.

Spatial features

Spatial surfaces are represented as a series of **features** which can be represented as multi-variate variables containing both the **spatial geometry** and **attributes**.

Spatial geometry describes where a feature is physically located in 2d (longitude, latitude) or 3d (longitude, latitude, altitude) space. This can be represented as different **geometric types** like a polygon, line, point, etc. What all geometric types have in common is that they are represented as a single or a series of **location coordinates**. Location coordinates can be represented in many ways but most commonly as 2d or 3d vectors.

The attribute represents the value of a measurement at a certain location coordinate and can be of any type. An example of a single 2d feature for a windspeed surface could be $(x_i = \{10.2, 23.2\}, y_i = 3.424)$.

Coordinate reference system

A major feature of any GIS is the coordinate reference system (CRS). It is a system that describes the relationship of spatial geometry to locations on the Earth its surface. It is an essential component of any GIS since it allows data from many sources to be combined and analyzed.

A CRS is defined by its origin, orientation, and scale. The place on the Earth's surface where the CRS is centered is referred to as the origin. The orientation specifies the orientation of the coordinate axes with respect to the Earth its surface. The scale establishes the link between map distances and distances on the Earth its surface.

In a GIS a CRS is represented as big string of text representing a wide variety of parameters used for the translation between spatial coordinates with respect to the Earth its surface.

"EPSG:28992 +proj=sterea +lat_0=52.1561605555556 +lon_0=5.38763888888889 +k=0.9999079 +x_0=155000 +y_0=463000 +ellps=bessel +towgs84=565.4171,50.3319,465.5524,1.9342,-1.6677,9.1019,4.0725 +units=m +no_defs +type=crs"

Fig 1.2.1: Example of a CRS string used by a GIS system.

The method to specify the precision of the coordinates in the CRS is referred to as accuracy which is denoted in meters.

Spatial interpolation methods

The following chapters will briefly cover how several interpolation methods work, what category they belong in, their strengths, weakness and use cases and how they differ from each other.

Global	**Local**	
Deterministic	**Deterministic**	**Stochastic**
Trend Surface (approx.)	Inverse Distance Weighted (exact)	Universal Kriging (approx.)
Radial Basis Function (approx.)		Ordinary Kriging (approx.)

Figure 1.3.1: Spatial interpolation methods with their classification.

Trend surface

A trend surface can be generated for a set of receptors variables to capture the global trend from several scattered observations. It is based on low-order polynomials of spatial locations and fits a polynomial surface to the receptor variables using least-squares regression with the observations. This approach produces a surface with the lowest variation in proportion to the observation variables. Trend surfaces are typically described using simple linear polynomials for continuous values.

As the order of the polynomial is increased, the surface being fitted becomes more complex. Depending on the data, a higher-order polynomial may not always produce a more accurate surface. This is because as the model complexity increases, the number of included inflection locations and variable extremes are also increased. If a smoother global trend is desired using a higher order polynomial model might not be suitable.

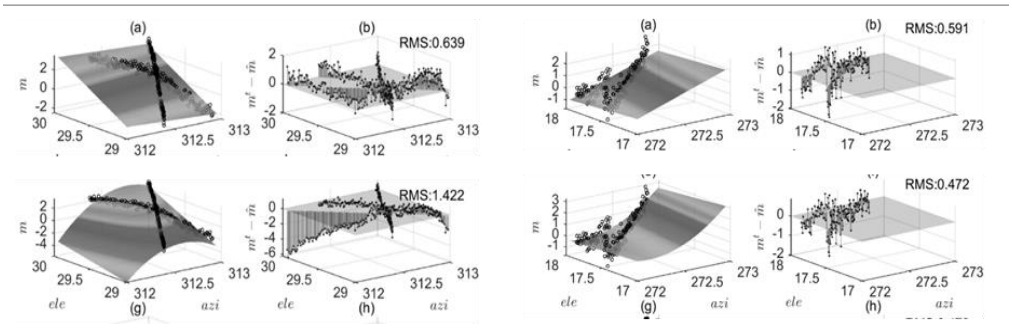

Figure 1.3.2: Example how data with a simple trend (left) results with a worse RMSE when using a higher order polynomial than with data containing a random trend (right).

This method is useful for comparisons with other interpolation methods to measure the performance gain from a baseline of the global trend of the data. Using a polynomial model with an excessively low order will not catch enough of the local fluctuations in the global trend, whilst using an order which is too high will not capture a generic enough global trend.

The receptor variable ($\hat{z}_i$) at receptor location (v_i) for a n^{th} order trend can be determined by summing up the values that result from a series of n order polynomial equations.

$$\hat{\boldsymbol{z}}_{\boldsymbol{i}}(\boldsymbol{v}_{\boldsymbol{i}}) = \boldsymbol{w}_{\boldsymbol{0}} + \boldsymbol{w}_{\boldsymbol{1}}\boldsymbol{v}_{\boldsymbol{ix}} + \boldsymbol{w}_{\boldsymbol{2}}\,\boldsymbol{v}_{\boldsymbol{iy}}$$

$$\boldsymbol{w}_{\boldsymbol{0}} := \boldsymbol{receptor}; \boldsymbol{w}_{\boldsymbol{n+1}}\ \boldsymbol{polynomial\ coefficients}$$

Fig 1.3.3: formula to calculate first order trend of 2d surface

Radial basis function

A radial basis function (RBF) is a mathematical function that calculates the distance between two locations in three dimensions. RBF interpolation is a non-parametric method that makes no assumptions about the underlying data distribution. The main idea behind RBF interpolation is to interpolate data using the sum of weighted radial basis-functions. These weights can be determined by minimizing the difference between the receptor variables using a sophisticated algorithm or neural network. These weights are normalized and typically a very low positive or negative continuous values within the range of values of the observations.

$$\min(z) < \sum_{j=1}^{n} w_j z_j < \max(z)$$

Fig 1.3.4: Normalized relationship of weights in RBF

RBF interpolation has many advantages over other methods of interpolation. It is computationally efficient when dealing with huge datasets with numerous variables. It also creates smooth, continuous, and distinguishable surfaces. It is, however, sensitive to the choice of basis- functions and necessitates careful parameter selection. Depending on the application this can have a considerable impact on the interpolation's accuracy. Gaussian, inverse multiquadric, and thin plate splines are some commonly radial basis functions.

The receptor variable ($\hat{z}_i$) at receptor location (v_i) can be calculated by summing the response of the chosen radial basis function $\phi(d)$ and its underlying radial basis weight of (w_i) several observations (n). The Euclidean distance ($\|...\|$) between the observation location (v_j) and receptor location (v_i) is used as the input of the radial basis function.

$$\hat{z}_i(v_i) = \sum_{j=1}^{n} w_j[\phi(\| v_i - v_j \|)]$$

Fig 1.3.5: Generalized formula of radial basis interpolation.

Inverse distance weighing

Inverse distance weighted interpolation (IDW) considers the degree of similarity and the power of neighboring samples in relation to distance. In a nutshell, the technique assumes that items that are close together are more similar than things that are far apart. IDW uses the observation variables surrounding a receptor to forecast the receptor variable. It is important to note, however, that IDW implicitly implies the presence of spatial autocorrelation in the data. When local variability is known to be present, this technique is appropriate to use.

It is a simple technique often implemented for meteorological data because of its simplicity. Here it is often used to compare performance gains with other kriging methods. (Zhao W. et al, 2022).

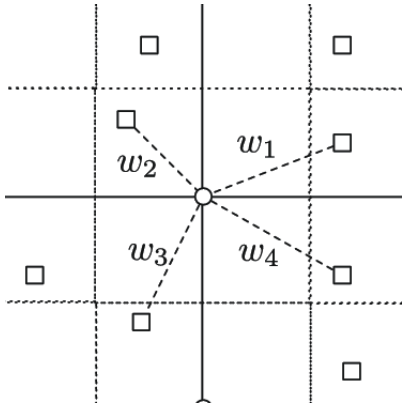

Fig 1.3.6: Example of IDW operation using the distance of the four closest neighbors.

The receptor variable ($\hat{z}_i$) at receptor location (v_i) can be calculated by using the distance (d_j) between the observation location (v_j) and receptor location (v_i), the observation variable (z_j) and by a user selected power parameter (w).

$$\hat{z}_i(v_i) = \frac{\sum_{j=1}^{n}[\frac{z_j}{d_j^w}]}{\sum_{j=1}^{n}[\frac{1}{d_j^w}]}$$

Fig 1.3.7: Simplified formula for inverse distance weighting interpolation.

In IDW the power parameter controls how rapidly the influence of an observation variable diminishes with distance from the receptor location. A power parameter close to one corresponds to linear interpolation, while a power parameter nearing two corresponds to inverse distance squared interpolation. It is a continuous, positive number. A bigger power parameter magnifies the effect of distance on the outcome.

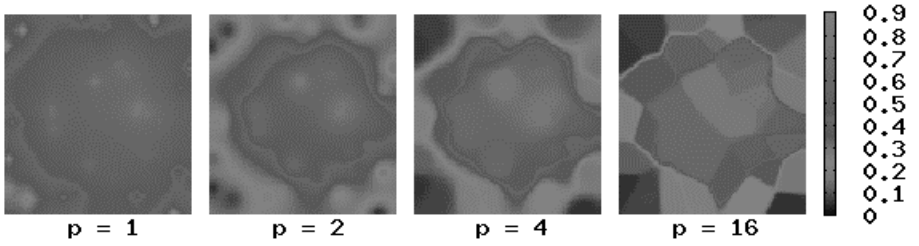

Fig 1.3.8: Results compared of a surface generated using inverse distance weighting interpolation with different power parameters.

Kriging

Kriging is a stochastic interpolation approach. It is predicated on the premise that the spatial correlation between features decreases with increasing distance between them. This geographical association is modeled using the semi-variance or covariance between two features.

There are a variety of different kriging methods available. Common ones used for interpolation of meteorological data are ordinary kriging; universal kriging and spatiotemporal kriging (Reinhardt et al 2018; Friedland C. et al 2017; Lee. C. 2022). Like IDW, kriging methods use a sample of observations close to a receptor to make predictions.

The most noteworthy distinction between these different kriging methods is the model that uses the semi-variances or covariances, whether it's suited for stationary or non-stationary data, whether the mean is constant or based on a model and whether auxiliary or temporal variables are used or not.

In the simplest form a kriging model (also referred to as simple kriging with semi-variance model) can be expressed as a weighted sum using semi-variances. The receptor variable ($\hat{z}_i$) at receptor location (v_i) can be calculated by using the kriging weights (w_j) of the (n) closest neighboring observations (z_j) including some error (ε_i).

$$\hat{\boldsymbol{z}}_{\boldsymbol{i}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} [\widehat{\boldsymbol{w}}_{\boldsymbol{j}} \boldsymbol{z}_{\boldsymbol{j}}] + \boldsymbol{\varepsilon}_{\boldsymbol{i}} \qquad\qquad \hat{\boldsymbol{z}} = \widehat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{z} + \boldsymbol{\varepsilon}$$

Fig 1.3.9: Simple kriging represented as a formula (left) and matrix form (right).

The matrix of kriging weights ($\hat{w}$) can be calculated by using a matrix of semi-variances **between each neighboring observations** around the receptor (D) and a matrix of semi-variances **between the receptor and each neighboring observation** (E).

$$\hat{w} = D^{-1}E$$

$$D := 2d\ matrix;\ E := 1d\ matrix;\ \hat{w} := 1d\ matrix$$

Fig 1.3.10: Matrix form of formula for calculation of kriging weights.

A semi-variance $\gamma(v_i, v_j)$ represents the variance of the difference between a pair of variables as a function of their location distance.

$$\gamma(v_a, v_b) = \frac{1}{2}(z_a - z_b)^2 \rightarrow z$$

Fig 1.3.11: Semi-variance formula between two features.

Calculating matrix (E) is rather straight-forward as you can just calculate the semi-variance between the receptor and the (n) closest neighbors.

$$E \rightarrow \{\gamma(\hat{v}_i, v_1) \ldots \gamma(\hat{v}_i, v_n)\}$$

Fig 1.3.12: Semi-variance matrix between receptor and neighboring observation.

Calculating matrix (D) requires you to calculate a matrix of semi-variance matrices $\gamma(h)$ for each set of lagged observations around the receptor. In each lag (matrix row) you pick different selection of pairs containing an equal amount of (k) pairs.

$$D \rightarrow \begin{pmatrix} \gamma(v_1, v_1) & \gamma(v_1, v_2) & \cdots & \gamma(v_1, v_k) \\ \gamma(v_2, v_1) & \gamma(v_2, v_2) & & \gamma(v_2, v_k) \\ \vdots & & \ddots & \vdots \\ \gamma(v_k, v_1) & \gamma(v_k, v_2) & \cdots & \gamma(v_k, v_k) \end{pmatrix}$$

Fig 1.3.13: Semi-variance matrix neighboring observation around receptor.

Nugget, sill, and range

The nugget effect, sill, and range are essential kriging metrics which influence model performance. A strong nugget effect implies that there is a lot of variability at short distances and that a model with a smaller range should be employed. A high sill suggests that there is a lot of variability over long distances and that a model with a wider range should be used. A small range can result in overfitting, whereas a big range can result in underfitting.

The nugget is the variance between two points at zero distance. The sill is the spatial correlation model's greatest variance. The range is the distance at which the spatial correlation disappears. Their calculation includes the semi-variance at various lags. The nugget effect is the difference between the observed and nugget variance. It represents the microscale variability that the model cannot explain. The sill is the model's variation across long distances. The range is the distance at which the spatial correlation disappears.

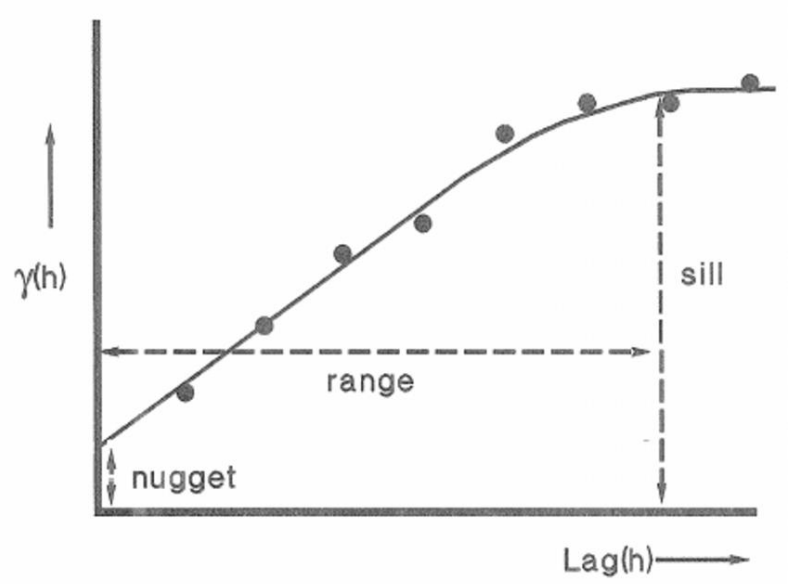

Fig 1.3.14: Example of semi-variogram showing the nugget, sill, and range, with $\gamma(h) \coloneqq semivariance$ *and* $Lag(h) \coloneqq lag\ distance$

Ordinary kriging

Ordinary kriging presumes that the mean of the receptor variables is constant and unknown and that the semi-variance between the variables at any two locations is only determined by the distance between the locations, hence assuming stationarity. It is the most basic and widely used kriging method.

The receptor variable ($\hat{z}_i$) at receptor location (v_i) can be calculated estimating the receptor mean ($\hat{\mu}_i$) and observation mean ($\hat{\mu}_j$) for the (n) neighboring observations, alongside using the observation variable (z_j) with the kriging weights (w_j) including some error (ε_i).

$$\hat{z}_i = \hat{\mu}_i + \sum_{j=1}^{n} w_j[(z_j - \hat{\mu}_j)] + \varepsilon_i$$

Fig 1.3.15: Simplified formula for ordinary kriging

Universal kriging

Universal kriging is an extension of ordinary kriging that presumes the mean is unknown and is modeled as a n^{th} order polynomial function or Gaussian model. To estimate the mean value at each location, a regression model is used.

For meteorological data spherical, exponential or gaussian models are typically used. When the spatial correlation declines fast with distance, a spherical model can be used. When the spatial correlation declines slowly with distance, an exponential model might be used. When the spatial correlation declines smoothly with distance, a Gaussian model can be used. (Burrough et al., 2015; Cressie, 2015).

Universal kriging uses the exact same formula as ordinary kriging, the only differences lay in the methods used to estimate the receptor and observation mean.

SLOO-CV cross validation

LOO-CV (leave-one-out cross-validation) is a model evaluation technique which divides the observations into two parts: training and testing. The training set includes all except one observation. This training set is used to generate a new model and the filtered observation is used as a receptor to be predicted. The predicted value is compared to the actual value of this test observation and this difference is stored in a list of residuals.

This process repeats for each individual observation until each observation has been used in the testing set at least once. The list of residuals is then used to calculate metrics like the MSE, RMSE and NRMSE.

SLOO-CV simply refers to LOO-CV using spatial features. This validation method is commonly used for spatial statistics when there is insufficient data to divide into training and testing sets or when there is no external validation set available to check the accuracy of predictions.

References

Li, Y., Min, D., Do, M. N., & Lu, J. (2016). Fast guided global interpolation for depth and motion. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9907 LNCS, pp. 717–733). Springer Verlag. https://doi.org/10.1007/978-3-319-46487-9_44

Scheuerer, M., Schaback, R., & Schlather, M. (2013, August). Interpolation of spatial data-A stochastic or a deterministic problem? European Journal of Applied Mathematics. https://doi.org/10.1017/S0956792513000016

Li, J., & Heap, A. D. (2014, March). Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling and Software. https://doi.org/10.1016/j.envsoft.2013.12.008

Zhao W., Zhong Y., Li Q., Li M., Liu J., Tang L. (2022). Comparison and correction of IDW based wind speed interpolation methods in urbanized Shenzhen. https://dx.doi.org/10.1007/s11707-021-0948-z

Reinhardt, K., & Samimi, C. (2018). Comparison of different wind data interpolation methods for a region with complex terrain in Central Asia. Climate Dynamics, 51(9–10), 3635–3652. https://doi.org/10.1007/s00382-018-4101-y

Friedland, C. J., Joyner, T. A., Massarra, C., Rohli, R. V., Treviño, A. M., Ghosh, S., … Weatherhead, M. (2017). Isotropic and anisotropic kriging approaches for interpolating surface-level wind speeds across large, geographically diverse regions. Geomatics, Natural Hazards and Risk, 8(2), 207–224. https://doi.org/10.1080/19475705.2016.1185749

Lee, C. (2022). Long-term wind speed interpolation using anisotropic regression kriging with regional heterogeneous terrain and solar insolation in the United States. Energy Reports, 8, 12–23. https://doi.org/10.1016/j.egyr.2021.11.285

Burrough, P.A., McDonnell, R.A., Lloyd, C.D., 2015. Principles of Geographical Information Systems. Oxford University Press. https://www.researchgate.net/publication/37419765_Principle_of_Geographic_Information_Systems

Cressie, N., 2015. Spatial Prediction and Kriging. (2015) (pp. 105–209). https://doi.org/10.1002/9781119115151.ch3