Lecture 13. Indicator Kriging

Math 586

Assumption of normality: important. Kriging may give unacceptable results if the data are severely non-normal.
Also, sometimes we need to find probabilities (e.g. contamination)

$$P(V(\mathbf{x}_0) \le a \,|\, V_1, ..., V_n)$$

Under normality, they can be found using kriging mean $= \mathbb{E}\,[V(\mathbf{x}_0) \,|\, V_1, ..., V_n]$ and variance $\sigma_K^2 = Var[V(\mathbf{x}_0) \,|\, V_1, ..., V_n]$. The normal probability is found using standard normal CDF

$$\Phi\left(\frac{a - \texttt{mean}}{\texttt{st.dev.}}\right)$$

Possible remedies for non-normality: transformation (either functional or "normal score transform", see Lecture 3.)

Indicator kriging: another alternative. Also can be used independently when the data are inherently of categorical nature, e.g. vegetation type, soil type, success/ failure in drilling etc. (More than two types may require co-kriging, though.)

Indicator function: 0 or 1 values. For a random variable V, may define

$$I_{(a,b]} = \begin{cases} 1 & \texttt{if} \quad a < V \le b \\ 0 & \texttt{otherwise} \end{cases}$$

$I_{(a,b]}$ is itself a random variable. We may replace the problem of predicting $V(\mathbf{x})$ by the problem of predicting $I_{(a_i,b_i]}$ for several intervals $(a_i, b_i]$.

What does the fractional value of kriging prediction $\hat{I}_{(a_i,b_i]}(\mathbf{x}_0)$ signify? It's the probability that $a < V(\mathbf{x}_0) \le b$.

The method proceeds as follows:

- Convert the given values to indicators, for chosen intervals $(a_i, b_i]$, usually chosen to divide the range of V evenly
- Estimate the *indicator variograms* for each range interval.

- Do the kriging for each range interval, using the usual equations and obtain predictions.

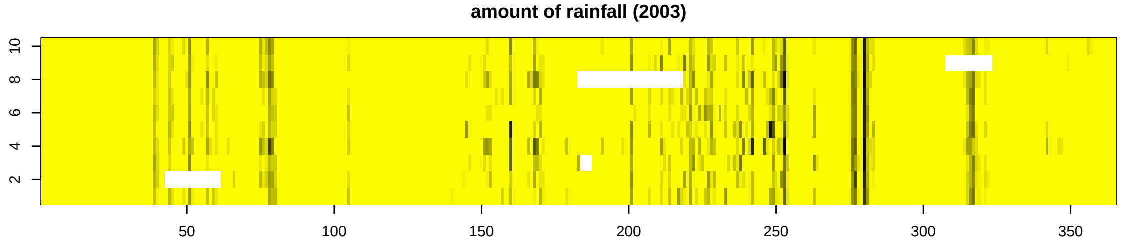
Difficulties:

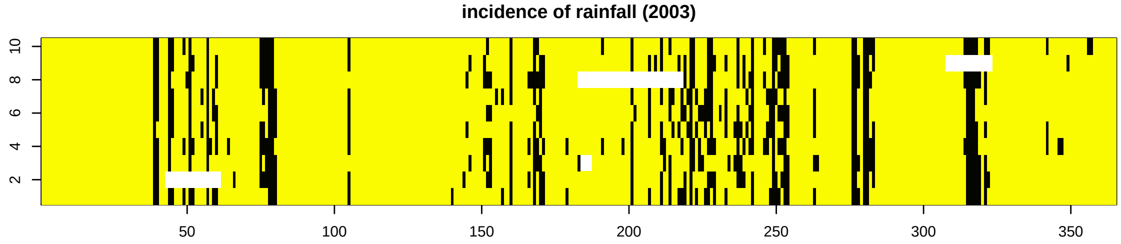
- will not necessarily get probabilities to add up to 1.
- sometimes the prediction may end up beyond interval $[0, 1]$ (e.g. kriging occasionally gives negative weights)
- loss of precision due to discretizing the range of V.

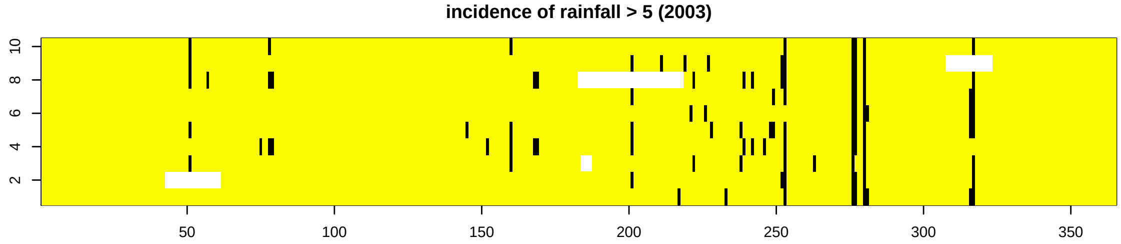
Example: rainfall prediction

We will use indicator kriging for interpolating the probability of rainfall on given days at Sevilleta. The indicator data we will consider are 0 if no rainfall and 1 if rainfall.

The data are collected for 10 stations in 2003. Some data are missing. The precipitation "map" with indicators is:







There are additional ideas that we will use: a) spatiotemporal modeling and b) moving window.

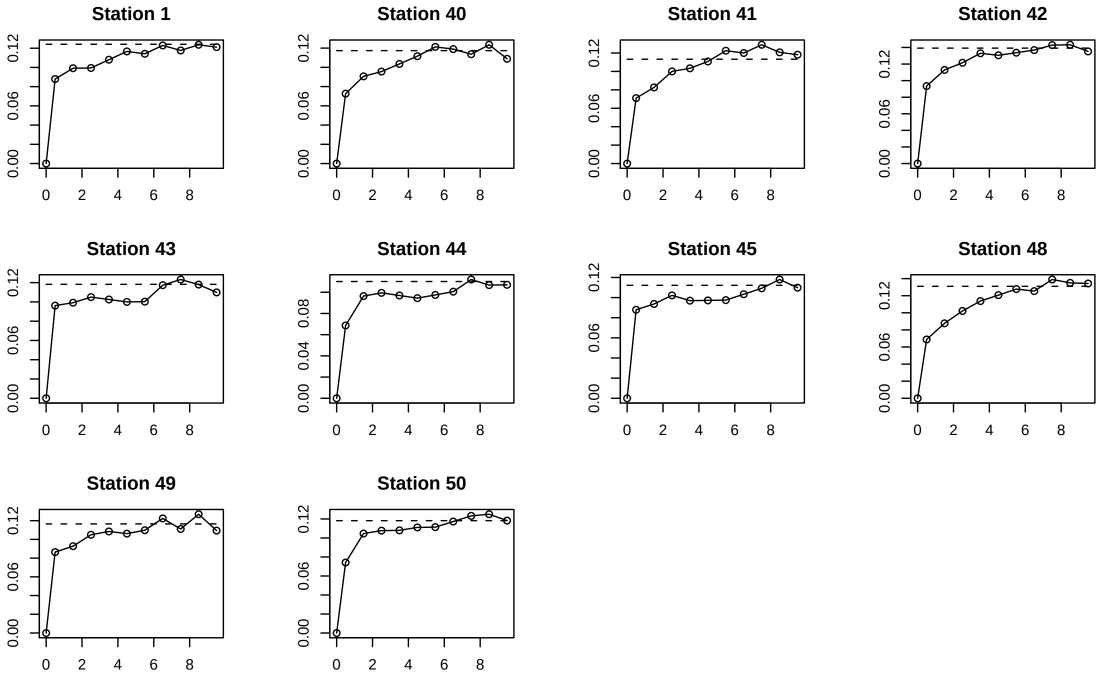
The moving time window will be used as the "local" option within Ordinary

Kriging. We can hardly expect the precipitation to be stationary throughout the year!

Step 1

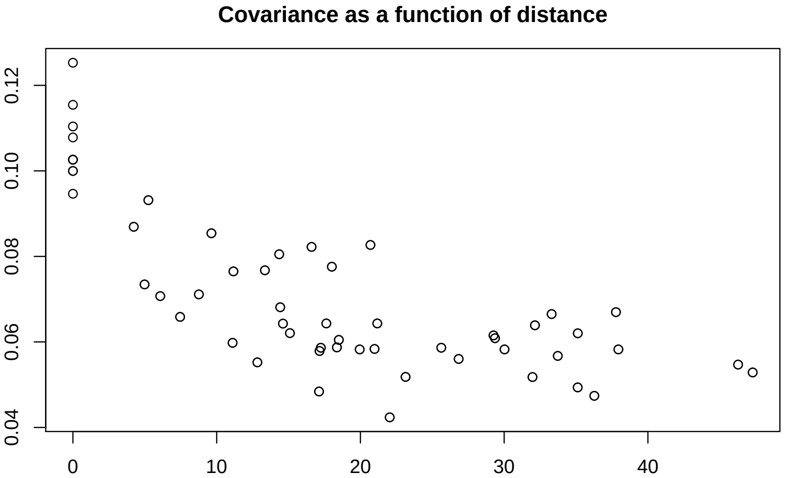
First, we will estimate space and time covariance functions. We will use a *separable* model, that is, fit the functions separately in space and time. See the code `SevIK.m` and data at `Sev03.txt`.

The time variograms were obtained for each station separately, and the results are



Note that for indicator (0-1) data the sill would be theoretically equal to Binomial variance $p(1-p)$, where p is the proportion of 1's in the sample. The above variograms appear consistent with exponential model and sill ≈ 0.11. For simplicity, we'll use temporal scale = 1 day. We don't expect a nugget.

The space variograms are easier estimated using the sample covariance between vectors of yearly observations for each pair of stations. Plotting those against the distance, we obtain



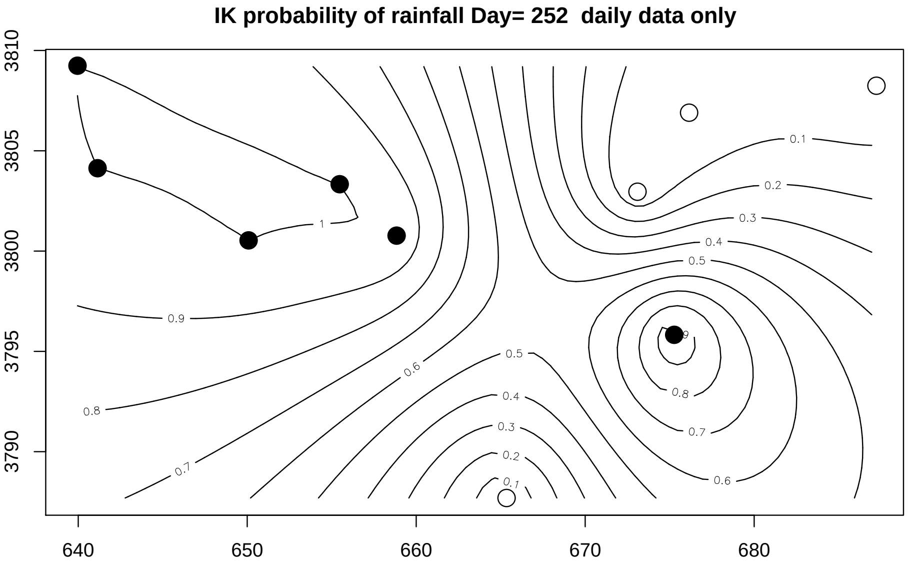
Again, the sill is near 0.11, which is encouraging, and we can surmise an exponential model with spatial scale ≈ 50km.

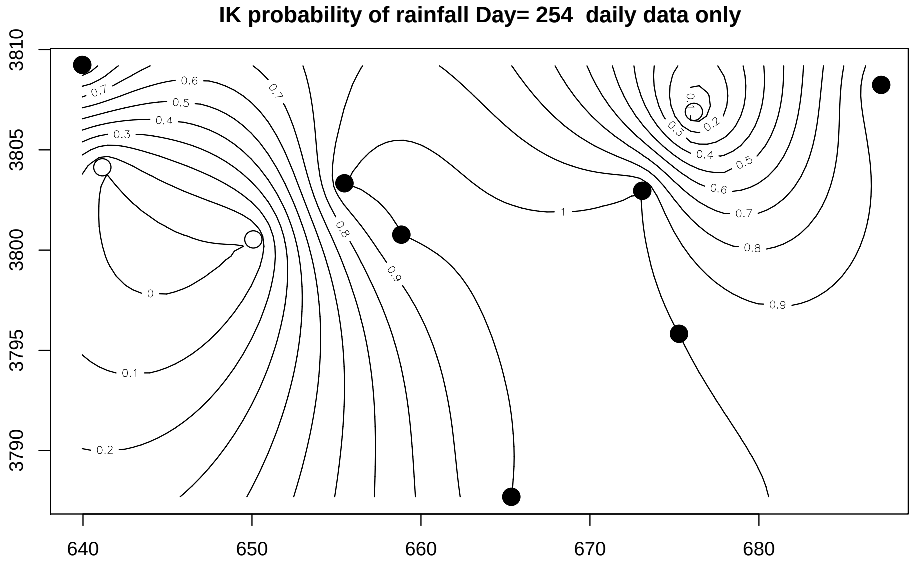
Step 2

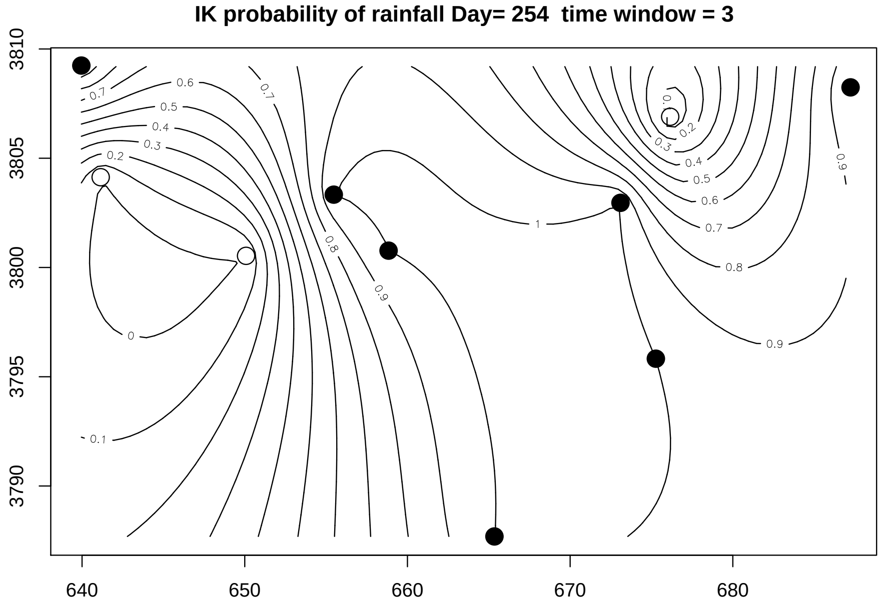
Now, we can compute the combined space-time covariance function for two locations $\mathbf{x}$ and times t:

$$Cov(\mathbf{x}_1, \mathbf{x}_2; t_1, t_2) = \sigma^2 e^{-|\mathbf{x}_1 - \mathbf{x}_2|/\texttt{SpaceScale}} \cdot e^{-|t_1 - t_2|/\texttt{TimeScale}} \tag{1}$$

To obtain a kriging map for the precipitation probability, use the data within a moving time-window $\pm w$: to obtain a map for day T, use all the observations (0-1 indicators) for the days $T - w, ..., T, ..., T + w$. The covariance matrix C is obtained using (1) for $|t_1 - t_2| \leq w$. Some results are below:




IK probability of rainfall Day= 254 daily data only


IK probability of rainfall Day= 254 time window = 3

Hard and Soft data

Sometimes, along with the *hard* data (direct observations), other information is available: expert opinion, prior guesses etc. These so-called *soft* data may be in the interval form (“I know that the porosity here is between 15 and 20%”) or in the form of a *prior* probability distribution (“I know that the porosity here follows normal distribution with the mean of 17% and st.dev. of 3%”). Both are potentially useful for kriging and can be incorporated via indicator kriging.

Introduce cutoffs $v_1, ..., v_k$.

- For interval data: inequality constraints $a_i < V(\mathbf{x}_{2i}) \leq b_i$. Then set k-th indicator $= 0$, if $v_k < a_i$, $=1$ if $v_k > b_i$ and undefined (no data) if $a_i < v_k \leq b_i$.

- For prior distirbution data of the form of cumulative distribution function $P(V(\mathbf{x}_{3i}) \leq v) \equiv F(v; \mathbf{x}_{3i})$ set k-th indicator simply equal to $F(v; \mathbf{x}_{3i})$

Use of different data types may require co-kriging.

More advanced Bayesian methods and conditional simulation may also be used.