About the Use of Shepards Interpolation Method by the Global Precipitation Climatology Centre

Jürgen Grieser

May 11, 2011

1 Motivation

The Global Precipitation Climatology Centre (GPCC) interpolates precipitation from $K \approx 3300$ stations in Germany to a 0.5 degree grid using Shepards Method in spherical coordinates by Spheremap (e.g. in their full-data product). The program was installed in the GPCC by Dr. David Legates about 20 years ago.

Here I show that this is not the most intelligent thing to do since method, data density and grid resolution do not fit to each other. When employed at the GPCC I therefore refused to use Shepards Method and asked to be allowed to interpolate to a finer grid. However, this was refused by my boss, Dr. Bruno Rudolf, head of the GPCC for more than a decade.

In order to keep things easy I use first order approximations. Thus the numbers may not be exact. The message, however, is clear.

The GPCC interpolates to a 0.5 degree grid. Thus a grid cell has the size

$$A_g = 2\pi R \frac{0.5}{360} \cdot 2\pi R \frac{0.5}{360} \cdot \cos(\varphi) \quad (1)$$

with the latitude φ and the earth radius $R = 6371km$. At 50 degrees latitude a grid cell has the size $A_g = 1987km^2$. Germany has a size of $A = 357 \cdot 10^3 km^2$ leading to $N = 179$ grid points in Germany.

2 Deterministic Approach

For the sake of simplicity we assume here that precipitation stations are exactly equidistantly distributed in Germany. This means that we can estimate the distance between two neighboring stations from the area $A_S = r_S^2 \pi$ that is represented by a station and the total area of Germany A. If there are K stations then the distance between 2 stations is $d \approx 2\sqrt{\frac{A}{K\pi}}$ where we have assumed that the area represented by a station is a perfect circle. We get similar results if we use hexagons instead which fill the area exactly. Shepards method is designed to produce the best point estimate for each location. Therefore, the method ignores all stations but those closer than $5km$ to a grid point in case that at least one station is that close to a grid point. Only one station per grid point is used if only one station is closer than $5km$ to the grid point. For equidistantly distributed stations this means a station distance of $10km$. And for the case of Germany this is realized by the use of

$$K* = \frac{A}{r_S^2 \pi} = \frac{357 \cdot 10^3 km^2}{25 * \pi} \approx 4545 \text{ stations.} \tag{2}$$

This means that in case that 4545 equidistantly distributed station in Germany were provided to Shepards Method only 179 would be used. The GPCC suggests (and does) interpolate by using Shepards Method with about 3300 stations. Therefore more stations are used for the interpolation. My suggestion to interpolate to a finer grid or use another interpolation method, however, was rejected by the GPCC in 2004.

3 Stochastic Approach

Now we assume that the German Met Service (DWD) distributed their stations not exactly equidistantly within Germany. However, in order to get a representative sample, station density is assumed to be the same everywhere. This means that we can expect to have $n = \frac{K}{N} = \frac{3300}{179} \approx 18.4$ stations per 0.5 degree grid cell. Within a grid cell stations may be randomly distributed. Since Shepards Method only uses a maximum of 10 neighboring stations, it would not use $\frac{18.4-10}{18.4} = \frac{8.4}{18.4} \approx 45.7\%$ of the stations (in numbers 1506) just because others are closer to the grid points.

Another feature of Shepards Method is that it uses only stations closer than 5km to the grid point if at least one station is that near. So what is the chance p to have at least one station closer than 5km to a grid point? For each station this is just the ratio p_1 of the area A_5 closer than $5km$ to the grid point to the area A_g attributed to a grid point. A_5 can be calculated easily to be $A_5 = r^2\pi = 78.54km^2$. Therefore $p_1 = \frac{A_5}{A_g} \approx 0.039 \approx 1/25$. Since we have n stations per grid cell the chance of having none within the area A_5 is $1 - p = (1 - p_1)^n$ which in case of n=18.4 yields 47.5%. This means that in more than 50% of the cases at least 1 station is closer to a grid point than $5km$. We can generalize this approach and ask how likely it is to have at least j stations closer than $5km$ to a grid point. The likelihood is just the product of having already $j - 1$ stations closer $5km$ times the likelihood of having at least one of the residual stations closer than $5km$ to the grid point and thus $p_j = p_{j-1} \cdot (1 - p_1)^{n-j}$. In the end we might expect using E_k stations with

$$E_k = (1 - p_1)^n \cdot N \cdot 10 - \sum_{j=1}^{n} (p_j - p_{j-1}) \cdot N \cdot j \tag{3}$$

where the first term is the fraction of grid points where 10 neighboring stations are used since none is closer than $5km$ to the grid point and the second term is the contribution of all grid points for which at least one and a maximum of 18 neighboring stations are used depending on how many fall closer than $5km$ to the grid point. For the case of 3300 stations in Germany we get an expected number of used stations of 1032, which is less than one third. The vector $p_{j-1} - pj$ is $(47.5, 27, 13.6, 6.5, 3, 1.3, 0.5, 0.2, 0.07, \cdots)/100$ and the expected number of grid points with $0, 1, 2, 3, 4, \cdots$ stations closer than $5km$ is $(86, 49, 24, 12, 5, 2, 1, < 1, \cdots)$. Please keep in mind that the use of even more stations would lead to less stations used until a minimum is reached when for the majority of grid points stations exist which are closer than $5km$.

Note also that as soon as at least one station is closer than 5km to a grid point only the area fraction $A_5/A_g \approx 1/25$ becomes sampled at that grid point. The area sampled A_σ is the sum of A_g times the fraction of grid points having no station closer than $5km$ plus $1/25A_g$ times the fraction of grid points having at least one station closer than $5km$ and thus

$$A_\sigma = A_g \cdot \left((1-p_1)^n + \frac{1}{25}(1-(1-p_1)^n)\right) \tag{4}$$

or as a relative sampling area

$$a_\sigma = (1-p_1)^n + \frac{1}{25}(1-(1-p_1)^n). \tag{5}$$

This converges against $1/25$ for a large number of stations. It is a monotonic function showing that the area sampled becomes smaller the more stations are used by using Shepards method for the interpolation of precipitation in Germany.

The more stations the GPCC provides to its interpolation method the less area is sampled. This should be kept in mind when using products of the GPCC.