Flexible Spatial Models on the Example of Temperature in China

Anastasija Teterева

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
Motivation

The objective of spatial interpolation is to create a continuous surface from a discrete set of points. Spatial prediction of weather phenomena are widely used in:

- **environmental science**;
- **industry** for planning;
- **ecology** to study greenhouse effect;
- **weather index-based insurance**.
Outline

1. Motivation
2. Data and Descriptives
3. Regression
4. Inverse distance weighting
5. Kriging
6. Copula-based interpolation
7. IDW-GEV interpolation
8. Conclusion
Data

Average temperature in **159 meteorological stations** in China over **53 years** (from January 1, 1957 till December 31, 2009). **Longitude, latitude** and **elevation** of each station are given.

- Average temperature is the average of max and min.
- No observations from Tibet (Xizang) and Jilin provinces.
- Weather stations in Xinjiang, Hunan and Neimongol provinces are widely spaced.
- 147 missing values were replaced.
Observed stations and climatic zones

Figure 1: Weather stations in China grouped by clusters and climatic zones.

Flexible Spatial Models on the Example of Temperature in China
Descriptive statistics 1

Figure 2: Temperature of 5 weather stations grouped by month.
### Descriptive statistics II

<table>
<thead>
<tr>
<th>Station</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaoyang</td>
<td>-22.90</td>
<td>-2.50</td>
<td>11.10</td>
<td>9.13</td>
<td>21.00</td>
<td>33.40</td>
<td>12.90</td>
</tr>
<tr>
<td>Dulan</td>
<td>-21.10</td>
<td>-4.90</td>
<td>3.80</td>
<td>3.20</td>
<td>11.30</td>
<td>25.60</td>
<td>9.34</td>
</tr>
<tr>
<td>Aershan</td>
<td>-40.50</td>
<td>-16.70</td>
<td>-0.20</td>
<td>-2.58</td>
<td>11.90</td>
<td>27.60</td>
<td>15.66</td>
</tr>
<tr>
<td>Haozhou</td>
<td>-11.90</td>
<td>5.80</td>
<td>15.90</td>
<td>14.88</td>
<td>23.90</td>
<td>34.70</td>
<td>10.02</td>
</tr>
<tr>
<td>Guilin</td>
<td>-2.90</td>
<td>12.30</td>
<td>20.20</td>
<td>19.00</td>
<td>26.10</td>
<td>33.00</td>
<td>7.86</td>
</tr>
</tbody>
</table>

**Table 1**: Numerical summary for 5 weather stations.

<table>
<thead>
<tr>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>30.88</td>
<td>976.05</td>
<td>1600.67</td>
<td>1683.13</td>
<td>2312.26</td>
<td>4480.88</td>
</tr>
</tbody>
</table>

**Table 2**: Numerical summary for distances between the stations.
Regression

Chuanyan et al. (2005) propose to model $Z_t(x_i)$ as linear function of the geographical characteristics $g_j(x_i)$:

$$Z_t(x_i) = \sum_{j=1}^{J} a_{t,j} \cdot g_j(x_i) + \varepsilon_t(x_i); \ t = 1, \ldots, T; \ i = 1, \ldots, 159.$$  

We use latitude, longitude and logarithm of elevation as $g_j(x_i)$.

Mean absolute error (out-of-sample) for station $i$:

$$\text{MAE}_i = \frac{1}{T} \sum_{t=1}^{T} |Z_t(x_i) - \hat{Z}_t(x_i)|$$

is evaluated using leave-one-out crossvalidation.
Regression error

- $R^2$ varies from 0.36 to 0.97
- $R^2$ strongly depends on the season
- Error does not "explode" in the mountain regions

**Figure 3:** MAE for regression model.
Inverse distance weighting (IDW)

The inverse distance interpolation formula is given by

\[ \hat{Z}_t(x_0) = \frac{\sum_{j: \|x_j - x_0\| \leq h} w(x_j) Z_t(x_j)}{\sum_{j: \|x_j - x_0\| \leq h} w(x_j)} \], \quad w(x_j) = \frac{1}{\|x_j - x_0\|^p} \]

We choose optimal \( p \) and \( h \) for each station

- \( h_i = \arg \min_{h \in [Q_{0.05}, Q_1]} \sum_{t=1}^{T} |Z_t(x_i) - \hat{Z}_t(x_i)| \)
- \( p_i = \arg \min_{p \in [0.5, 20]} \sum_{t=1}^{T} |Z_t(x_i) - \hat{Z}_t(x_i)| \)
- \( Q \) is empirical quantile of distances between the stations
Choosing $p$ and $d$

Figure 4: Optimal $p$ (left) and $h$ (right) for station $i = 26$. 
Choosing $p$ and $d$

Figure 5: Optimal $p$ and $h$ for each station.
IDW interpolation error

- IDW error strongly depends on $p$ and $d$
- There is no spatial pattern in $p$ and $d$
- We choose $p = 3$ and $d = 556$ minimizing MAE over all stations
- MAE strongly depends on region

Figure 6: MAE for IDW model.
Universal kriging

The empirical variogram is given by

\[ 2\hat{\gamma}_n(h) = \frac{1}{\# N(h)} \sum_{(x_i, x_j) \in N(h)} \{Z(x_i) - Z(x_j)\}^2, \ h \in \mathbb{R}^r. \]

\[ N(h) = (x_i, x_j) : (r - \delta) \leq \|x_i - x_j\| \leq (r + \delta); \ i, j = 1, \ldots, n, \ r = \|h\| > 0. \]

We use Gaussian model \( \gamma(h) = c + (s - c) \left(1 - \exp \frac{-3h^2}{a^2}\right) \) and calculate the weights according to

\[
\begin{bmatrix}
\lambda_1 \\
\vdots \\
\lambda_n \\
\mu
\end{bmatrix} = \begin{bmatrix}
0 & \cdots & \hat{\gamma}(x_1, x_n) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\hat{\gamma}(x_n, x_1) & \cdots & \hat{\gamma}(x_n, x_n) & 1 \\
1 & \cdots & \cdots & 1
\end{bmatrix}^{-1} \begin{bmatrix}
\hat{\gamma}(x_1, x_0) \\
\vdots \\
\hat{\gamma}(x_n, x_0)
\end{bmatrix}
\]
Fitting the variogram

Data are anisotropic with two main directions:

\[ \alpha = 0 \]

\[ \alpha = \pi/2 \]

**Figure 7**: Directional empirical variograms and fitted Gaussian models.
Kriging interpolation error

- Kriging gives similar to IDW error structure
- MAE is region dependent
- Error is smaller in the coastal area and larger in the mountain areas

Figure 8: MAE for kriging model.
Copula-based interpolation

Kazianka (2010) and Bardossy (2011) propose to model dependence of any two locations separated by the vector $h$ by

$$P\{Z(x_i) \leq z_i, Z(x_j) \leq z_j\} = C_h\{F_Z(z_i), F_Z(z_j)\}.$$  

They use the bivariate spatial copula

$$c_h(u, v) = \begin{cases} 
  c_{1,\tau(h)}(u, v) & \text{, if } 0 \leq h < l_1 \\
  (1 - \lambda_2)c_{1,\tau(h)}(u, v) + \lambda_2 c_{2,\tau(h)}(u, v) & \text{, if } l_1 \leq h < l_2 \\
  \vdots & \text{, if } l_{k-1} \leq h < l_k \\
  (1 - \lambda_k)c_{k-1,\tau(h)}(u, v) + \lambda_k & \text{, if } l_{k-1} \leq h \leq l_k \\
  1 & \text{, if } l_k \leq h 
\end{cases}$$

$$\lambda_j = \frac{h - l_{j-1}}{l_j - l_{j-1}}.$$  

We propose to choose copula and model its parameters as a function of distance and angle.
Copula-based interpolation algorithm

- Estimate marginals
  - Estimate GEV parameters for each station and each day of the year (e.g. for station $i = 26$ and $d = 18$th of July)
  - Model dependence of GEV parameters from geographical coordinates (use multiple linear regression)

- Estimate copula family
  - Choose bivariate copula
  - Estimate copula parameter for each pair of stations
  - Model copula parameter as function of separating distance $h$ and angle $\alpha$
Checking data for serial dependence

Figure 9: ACF (left) and PACF (right) of temperature for $i = 26$ and $d = 18$th of July. ADF test $p$-value $< 0.01$, Ljung-Box test $p$-value $= 0.61$. 

Flexible Spatial Models on the Example of Temperature in China
Checking data for serial dependence

Figure 10: ACF (left) and PACF (right) of squared temperature for $i = 26$ and $d = (t \mod 365) = 18$th of July.
Assessing the quality of a fitted GEV

**Figure 11:** Goodness of fit for GEV distribution ($i = 26$ and $d = 18$th of July).

Flexible Spatial Models on the Example of Temperature in China
Modeling GEV parameters - $\mu$

\[ \mu_d(x_i) = \sum_{j=0}^{2} a_{\mu,d,j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\mu,d,j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\mu,d,j} \log\{\text{El}(x_i)\}^j + \varepsilon_d(x_i) \]

Figure 12: $\mu_{200}$ as nonparametric and multiple linear regression of Lat, Lon and log(El).

- The chosen model is

Flexible Spatial Models on the Example of Temperature in China ———
Modeling GEV parameters - $\sigma$

Figure 13: $\sigma_{200}$ as nonparametric and multiple linear regression of Lat, Lon and log(El).

The chosen model is

$$
\sigma_d(x_i) = \sum_{j=0}^{3} a_{\sigma, d, j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\sigma, d, j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\sigma, d, j} \log\{\text{El}(x_i)^j\} + \varepsilon_d(x_i)
$$
Modeling GEV parameters - $\xi$

**Figure 14:** $\xi_{200}$ as nonparametric and multiple linear regression of Lat, Lon and log(El).

- The chosen model is

$$\xi_d(x_i) = \sum_{j=0}^{4} a_{\xi,d,j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\xi,d,j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\xi,d,j} \log\{\text{El}(x_i)\}^j + \varepsilon_d(x_i)$$
Choosing copula family

Figure 15: Contour plots suggest to choose Frank or elliptical family’s copula.
Modeling parameter of Gaussian copula

Figure 16: Gaussian copula parameter as nonparametric and multiple linear regression on separating distance \( (h) \), angle \( (\alpha) \) and logarithm of elevation difference \( \log\{\Delta(\text{El})\} \).

☐ The chosen model is

\[
\rho_d = \sum_{j=0}^{2} a_{\rho, d, j} h^j + \sum_{j=1}^{3} b_{\rho, d, j} \alpha^j + \sum_{j=1}^{3} c_{\rho, d, j} \log\{\Delta(\text{El})\}^j + \varepsilon_d
\]
Copula interpolation model (summary)

\[
\hat{Z}_t(x_0) = \int_0^1 F^{-1}_{\hat{\mu}_d(x_0), \hat{\sigma}_d(x_0), \hat{\xi}_d(x_0)} \{u(x_0)\} c_{\hat{\rho}_d} \{u(x_0) | Z_t(x_k)\} \, du(x_0)
\]

\[
\hat{\rho}_d = \sum_{j=0}^{2} a_{\rho,d,j} h^j + \sum_{j=1}^{3} b_{\rho,d,j} \alpha^j + \sum_{j=1}^{3} c_{\rho,d,j} \log \{\Delta(El)\}^j + \varepsilon_d
\]

\[
\hat{\mu}_d(x_i) = \sum_{j=0}^{2} a_{\mu,d,j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\mu,d,j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\mu,d,j} \log \{El(x_i)\}^j + \varepsilon_d(x_i)
\]

\[
\hat{\sigma}_d(x_i) = \sum_{j=0}^{3} a_{\sigma,d,j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\sigma,d,j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\sigma,d,j} \log \{El(x_i)\}^j + \varepsilon_d(x_i)
\]

\[
\hat{\xi}_d(x_i) = \sum_{j=0}^{4} a_{\xi,d,j} \text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\xi,d,j} \text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\xi,d,j} \log \{El(x_i)\}^j + \varepsilon_d(x_i)
\]
Copula interpolation error

- Error variation for different types of copulas is very small
- Copula-based interpolation reduces error in the mountain areas
- Gives larger error in the coastal area
- Is too complicated

Figure 17: MAE for copula model.
Do we really need copula?

$u_t(x_i) = \left\lfloor \frac{\text{rank}\{Z_\tau(x_i)\}}{54} \right\rfloor (t \text{ div } 365)$

$\tau = (d, d + 365, \ldots, d + 365 \cdot 52),
\quad d = (t \mod 365)$

$u_t(x_i)$ are grouped in clusters

We propose to estimate $u_t(x_0)$ with IDW and apply GEV quantile function to predict the temperature in unknown location

Figure 18: $u_t(x_i)$ pattern at $t = 200$. 
Simplified model

\[
\hat{Z}_t(x_0) = F^{-1}_{\hat{\mu}_d(x_0), \hat{\sigma}_d(x_0), \hat{\xi}_d(x_0)} \{\hat{u}_t(x_0)\}
\]

- \( \hat{u}_t(x_0) = \sum_{i: \|x_j - x_0\| \leq d} w(x_j) u_t(x_j) / \sum_{i: \|x_j - x_0\| \leq d} w(x_j) \)

- \( w(x_j) = 1 / \|x_j - x_0\|^p \)

- \( u_t(x_i) = [\text{rank}\{Z_t(x_i)\}/54](t \mod 365) \)

- \( \hat{\mu}_d(x_i), \hat{\sigma}_d(x_i), \hat{\xi}_d(x_i) \) as in copula interpolation formula

Flexible Spatial Models on the Example of Temperature in China
**Simplified model interpolation error**

IDW-GEW model gives small interpolation error in the coastal area and is able to capture extreme observations.

**Figure 19:** MAE for IDW-GEW model.
Comparing IDW and IDW-GEV models

- GEW-IDW model gives improvement for about 50% of the stations (number is season dependent)
- Improvement has no strong dependence on geographical coordinates
- GEW-IDW model is useful in prediction extremely high (low) temperatures

Figure 20: $(\text{MAE}_{\text{IDW}} - \text{MAE}_{\text{IDW-GEV}})$ for all stations at $t = 200$. 
Comparing IDW and IDW-GEV models

Figure 21: IDW (left) and IDW-GEV (right) prediction for station $i = 139$. 
Seasonal variation of error

Figure 22: MAE for regression, inverse distance weighting, kriging and EDW-GEV model for \( d = 1, \ldots, 365 \).

- All models give season dependent error
- IDW and IDW-GEV models are more robust
- IDW outperforms IDW-GEV model during the winter period
Conclusions

- The climate of China is extremely diverse - flexible interpolation techniques should be used
- Interpolation errors are region and time dependent for all discussed methods
- IDW, IDW-GEV and kriging give the smallest interpolation error
- Regression, copula-based and IDW-GEV interpolation are more robust methods
- IDW-GEV interpolation may be useful to handle extreme temperatures
References and articles:


Flexible Spatial Models on the Example of Temperature in China
References and articles:

- Gräler B., Kazianka H., de Espindola G. M. (2010): *Copulas, a novel approach to model spatial and spatio-temporal dependence*
References and articles:

Flexible Spatial Models on the Example of Temperature in China