Flexible Spatial Models on the Example of Temperature in China

Anastasija Tetereva

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

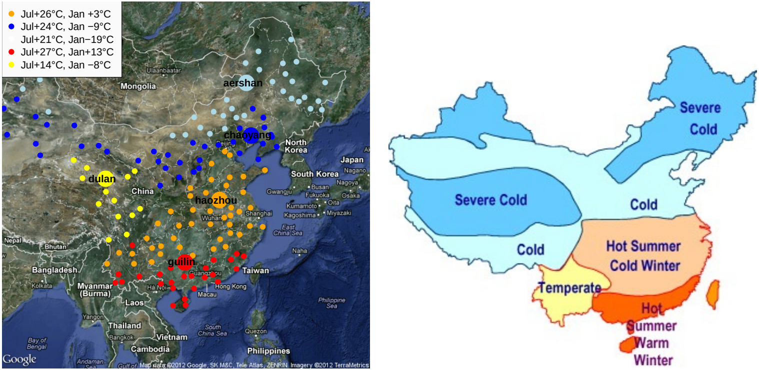


Figure 1: Weather stations in China grouped by clusters and climatic zones.

Descriptive statistics I

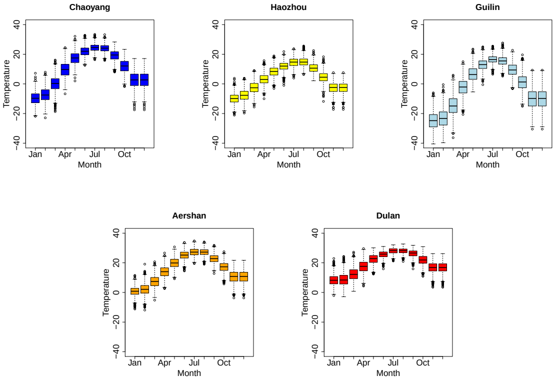


Figure 2: Temperature of 5 weather stations grouped by month.

Descriptive statistics II

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

Table 1: Numerical summary for 5 weather stations.

	Min	Q1	Median	Mean	Q3	Max	SD
distance	30.88	976.05	1600.67	1683.13	2312.26	4480.88	887.42

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 $g_j(x_i)$:

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

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

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

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

is evaluated using leave-one-out crossvalidation.

Regression error

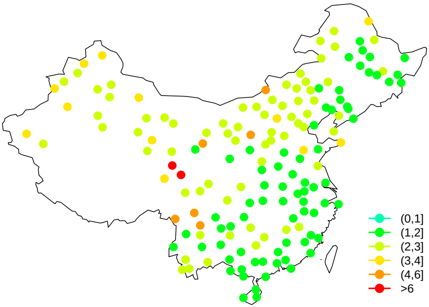


Figure 3: MAE for regression model.

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

Inverse distance weighting (IDW)

- The inverse distance interpolation formula is given by

$$\widehat{Z}_t(x_0) = \frac{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j) Z_t(x_j)}{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j)}, \quad w(x_j) = 1/\|x_j - x_0\|^p$$

,

- We choose optimal p and h for each station
  - $h_i = \arg \min\limits_{h \in [Q_{0.05}, Q_1]} \Sigma_{t=1}^{T} |Z_t(x_i) - \widehat{Z}_t(x_i)|$
  - $p_i = \arg \min\limits_{p \in [0.5, 20]} \Sigma_{t=1}^{T} |Z_t(x_i) - \widehat{Z}_t(x_i)|$
  - Q is empirical quantile of distances between the stations

Choosing p and d

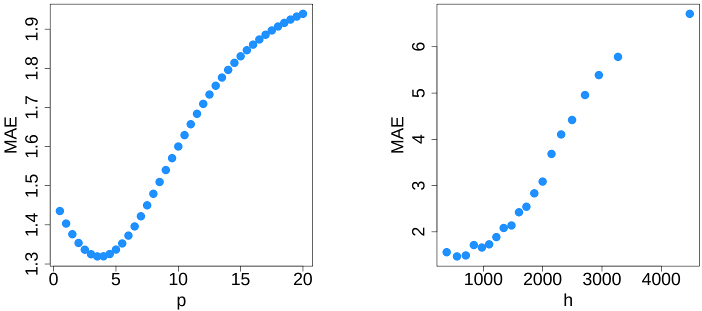


Figure 4: Optimal p (left) and h (right) for station $i = 26$.

Choosing p and d

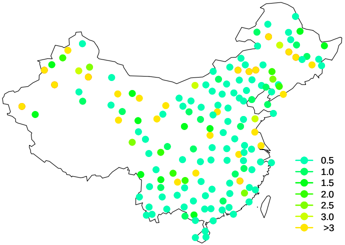


Figure 5: Optimal p and h for each station.

IDW interpolation error

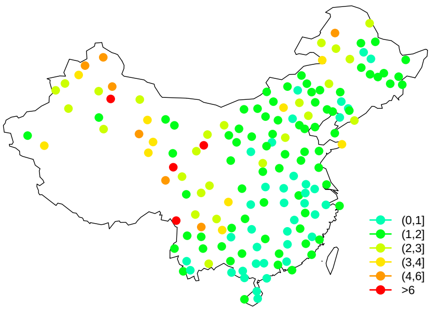


Figure 6: MAE for IDW model.

- 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 or region

Universal kriging

The empirical variogram is given by

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

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

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

$$\begin{bmatrix} \lambda_1 \\ \cdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \widehat{\gamma}(x_1, x_n) & 1 & \\ \cdots & \ddots & \cdots & \cdots & \\ \widehat{\gamma}(x_n, x_1) & \cdots & \widehat{\gamma}(x_n, x_n) & 1 & \\ 1 & \cdots & \cdots & 1 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} \widehat{\gamma}(x_1, x_0) \\ \vdots \\ \widehat{\gamma}(x_n, x_0) \\ 1 \end{bmatrix}$$

Fitting the variogram

Data are anisotropic with two main directions:

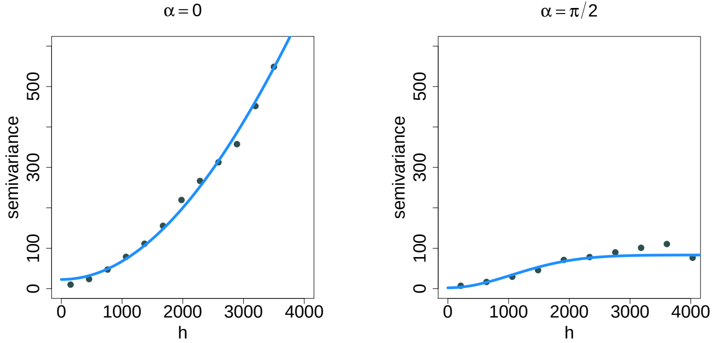


Figure 7: Directional empirical variograms and fitted Gaussian models

Kriging interpolation error

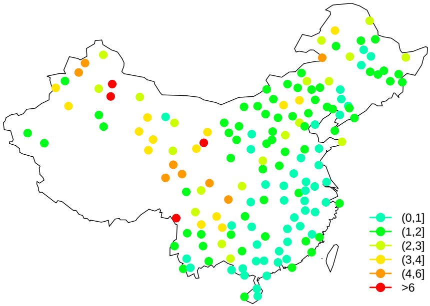


Figure 8: MAE for kriging model.

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

Copula-based interpolation

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

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

They use the bivariate spatial copula

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

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

Checking data for serial dependence

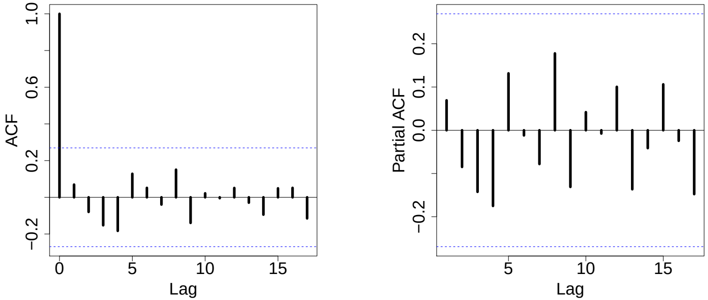


Figure 9: ACF (left) and PACF (right) of temperature for $i = 26$ and $d =$ 18th of July. ADF test p-value < 0.01, Ljung-Box test p-value $= 0.61$.

Checking data for serial dependence

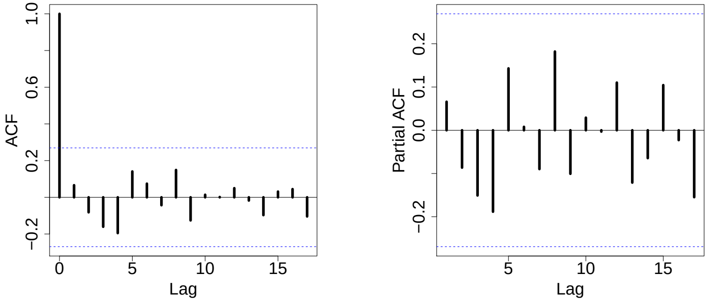


Figure 10: ACF (left) and PACF (right) of squared temperature for $i = 26$ and $d = (t \bmod 365) = $ 18th of July.

Assessing the quality of a fitted GEV

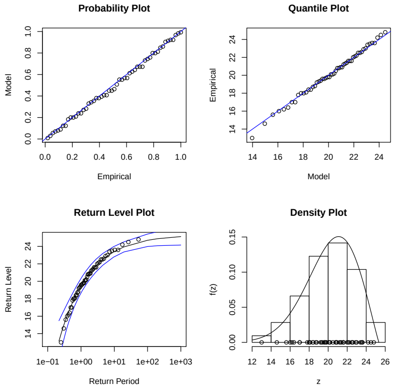


Figure 11: Goodness of fit for GEV distribution ($i = 26$ and $d = 18$th of July).

Modeling GEV parameters - μ

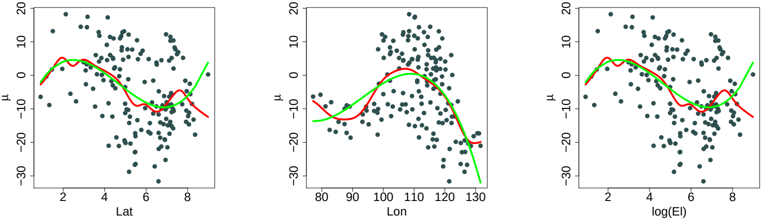


Figure 12: μ_{200} as nonparametric and multiple linear regression of Lat, Lon and log(El).

- The chosen model is

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

Modeling GEV parameters - σ

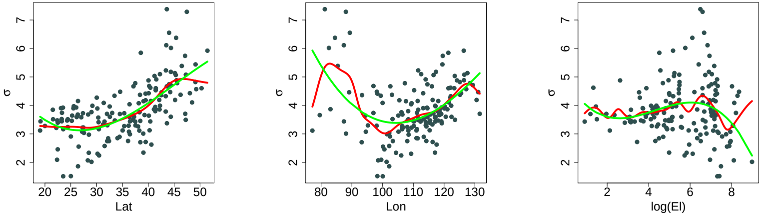


Figure 13: σ_{200} as nonparametric and multiple linear regression of Lat, Lon and log(El).

- The chosen model is

$$\sigma_d(x_i) = \sum_{j=0}^{3} a_{\sigma,d,j} \mathrm{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\sigma,d,j} \mathrm{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\sigma,d,j} \log\{\mathrm{El}(x_i)\}^j + \varepsilon_d(x_i)$$

Modeling GEV parameters - ξ

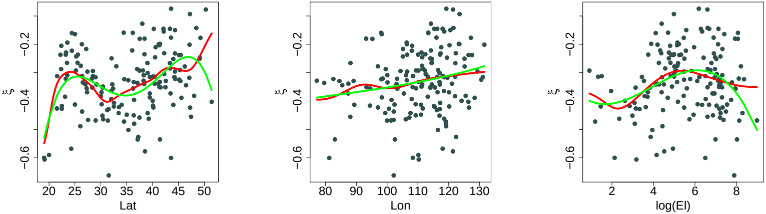


Figure 14: ξ_{200} as nonparametric and multiple linear regression of Lat, Lon and log(El).

- The chosen model is

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

Choosing copula family

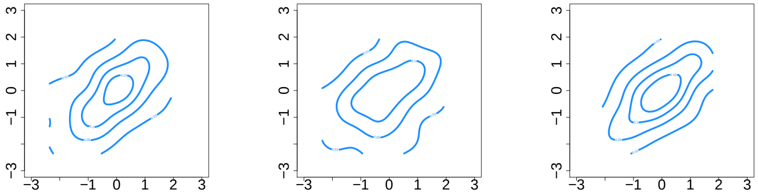


Figure 15: Contour plots suggest to choose Frank or elliptical family's copula.

Modeling parameter of Gaussian copula

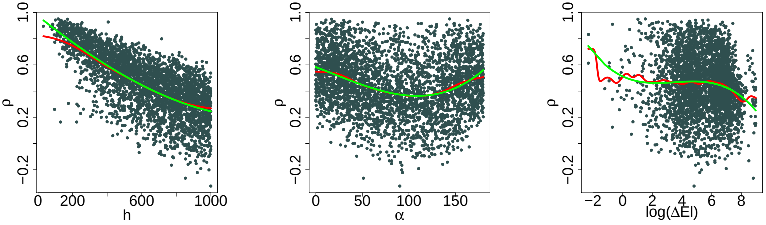


Figure 16: Gaussian copula parameter as nonparametric and multiple linear regression on separating distance (h), angle (α) and logarithm of elevation difference $\log\{\Delta(\mathrm{El})\}$.

- The chosen model is

$$\rho_d = \sum_{j=0}^{2} a_{\rho,d,j} h^j + \sum_{j=1}^{3} b_{\rho,d,j} \alpha^j + \sum_{j=1}^{3} c_{\rho,d,j} \log\{\Delta(\mathrm{El})\}^j + \varepsilon_d$$

Copula interpolation model (summary)

$$\widehat{Z}_t(x_0) = \int_0^1 F^{-1}_{\widehat{\mu}_d(x_0),\widehat{\sigma}_d(x_0),\widehat{\xi}_d(x_0)}\{u(x_0)\} c_{\widehat{\rho}_d}\{u(x_0)|Z_t(x_k)\}\,\mathrm{d}u(x_0)$$

$$\rho_d = \sum_{j=0}^{2} a_{\rho,d,j} h^j + \sum_{j=1}^{3} b_{\rho,d,j} \alpha^j + \sum_{j=1}^{3} c_{\rho,d,j} \log\{\Delta(\mathrm{El})\}^j + \varepsilon_d$$

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

$$\sigma_d(x_i) = \sum_{j=0}^{3} a_{\sigma,d,j} \mathrm{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\sigma,d,j} \mathrm{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\sigma,d,j} \log\{\mathrm{El}(x_i)\}^j + \varepsilon_d(x_i)$$

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

Copula interpolation error

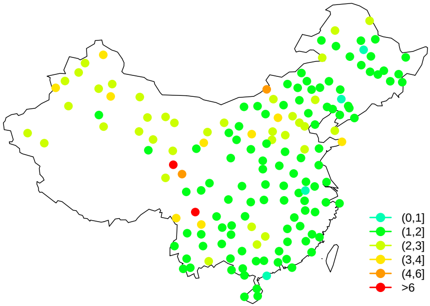


Figure 17: MAE for copula model.

- 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

Do we really need copula?

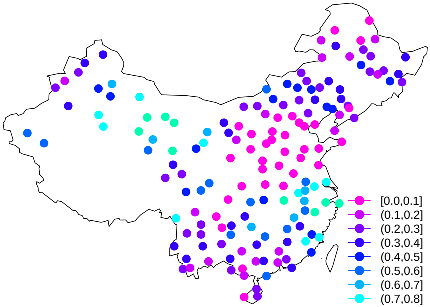


Figure 18: $u_t(x_i)$ pattern at $t = 200$.

- $u_t(x_i) = [\text{rank}\{Z_\tau(x_i)\}/54]_{(t \text{ div } 365)}$ $\tau = (d, d+365, \ldots, d+365 \cdot 52)$, $d = (t \bmod 365)$
- $u_t(x_i)$ 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

Simplified model

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

- $\widehat{u}_t(x_0) = \sum_{i:\|x_j - x_0\| \leqslant d} w(x_j) u_t(x_j) \; / \sum_{i:\|x_j - x_0\| \leqslant d} w(x_j)$

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

- $u_t(x_i) = [\text{rank}\{Z_\tau(x_i)\}/54]_{(t \text{ div } 365)}$

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

Simplified model interpolation error

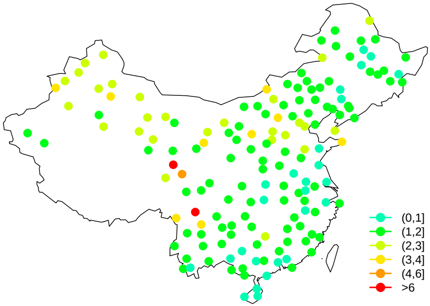
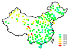


Figure 19: MAE for IDW-GEW model.

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



Comparing IDW and IDW-GEV models

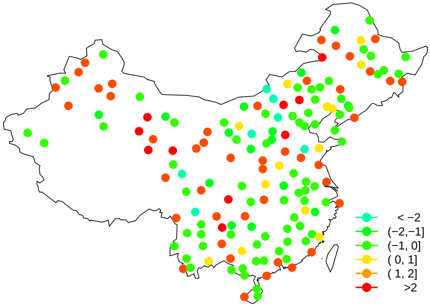


Figure 20: (MAE_{IDW}-$MAE_{IDW\text{-}GEV}$) for all stations at $t = 200$.

- 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

Comparing IDW and IDW-GEV models

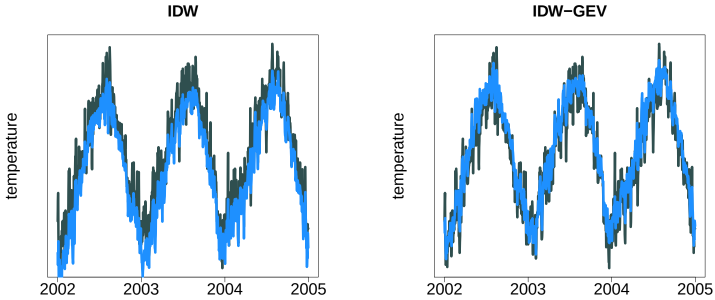


Figure 21: IDW (left) and IDW-GEV (right) prediction for station $i = 139$.

Seasonal variation of error

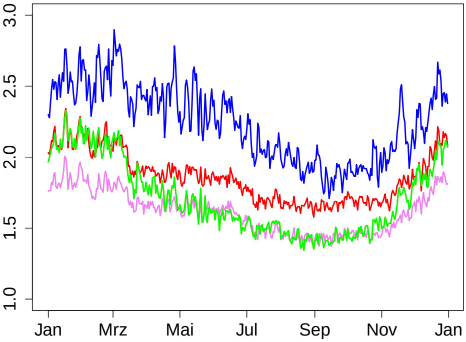


Figure 22: MAE for regression, inverse distance weighting, kriging and EDW-GEV model for $d = 1, \ldots, 365$.

- 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:

- Bardossy A. (2011): *Interpolation of Groundwater Quality Parameters with Some Values below the Detection Limit*, Hydrology and Earth System Sciences
- Chai H., Cheng W., Zhou C., Chen X., Ma X., Zhao S. (2002): *Analysis and comparison of spatial interpolation methods for temperature data in Xinjiang Uygur Autonomous Region, China*, Natural science
- Cressie N.A.C. (1991): *Statistics for Spatial Data*, John Wiley & Sons
- Diggle P.J., Ribeiro P.J. (2007): *Model-based Geostatistics*, Springer

References and articles:

- Gaetan C., Guyo X. (2010): *Spatial Statistics and Modeling*, Springer
- Gräler B., Kazianka H., de Espindola G. M. (2010): *Copulas, a novel approach to model spatial and spatio-temporal dependence*
- Härdle Wolfgang K.,López Cabrera B., Okhrin O., Wang W.(2011): *Localising temperature risk*, SFB 649 Discussion Papers
- Kazianka H., Pilz J. (2010): *Spatial Interpolation Using Copula-Based Geostatistical Models*, Quantitative Geology and Geostatistics

References and articles:

- Lauren H., Viger R., McCABE G. (2004): *Precipitation interpolation in mountainous regions using multiple linear regression*, Hydrology, Water Resources and Ecology in Headwaters
- Loecher M (2010): *Spatio-temporal analysis and interpolation of PM10 measurements in Europe*, ETC/ACM Technical Paper 2011/10
- Nelsen, Roger B. (2006): *An Introduction to Copulas*, Springer
- Pebesma E., Cornford D., Dubois G., Heuvelink G.B.M., Hristopoulos D., Pilz J., Stohlker U., Morin G., Skoien J.O. (2011): *INTAMAP: the design and implementation of an interoperable automated interpolation web service*, Computers & Geosciences