Construction of Spatial local C-vine model and application of air temperature interpolation

Lin Dai*, Jianxiu Liu

1 College of Science, Kunming University of Science and Technology, Kunming, Yunnan, 650000, China

*Corresponding author: 494816764@qq.com

Abstract. Based on the daily average temperature of 73 meteorological stations and the geographical location information of each station from 1980 to 2010 in Yunnan Province, spatial interpolation method of temperature data suitable for Yunnan province was explored. Based on the definition of spatial local C-Vine, the full model, horizontal distance model and vertical distance model were constructed. Using maximum likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) select the three models to obtain full model as the optimal model. 6 stations in the region were randomly selected and temperature interpolation was made using full model, Ordinary Kriging and Inverse Distance Weighting method. The results showed that full model based on Spatial local C-vine had better prediction accuracy.

1. Introduction
Air temperature data is mainly obtained through observations of weather stations. However, due to the limitation of economic cost, the number of meteorological stations is limited. It is difficult to obtain temperature data in remote areas. Scholars usually use spatial interpolation to solve this problem. Commonly used interpolation methods at home and abroad are Ordinary Kriging method (OK) and Inverse Distance Weighting method (IDW) [1-3]. Copula is a connection function that connects the marginal distribution of random variables to construct a multivariate composite distribution [4]. Bedford and Cooke [5] proposed Canonical vine(C-vine) and Drawable vine(D-vine). Copula theory has been widely used in the meteorology field. For example, Li Yan and others discussed the applicability of the Copula function in the study of spatial downscaling of surface temperature [6]. Leng Menghui et al. used Copula functions to carry out meteorological division of sponge cities [7]. Ye Minghua et al. used the binary t-Copula function to fit the dependence between temperature and precipitation [8]. For more related research results, please see Erhardt TM, Kazianka H, etc. [9].

Based on Spatial local C-Vine model and using the latitude, longitude, and altitude information of 73 meteorological observation stations in Yunnan Province, we used Spatial local C-vine established a horizontal distance model, a vertical distance model, and a full model. Combined with the daily average temperature data set of each station from 1980 to 2010, using maximum likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC) obtained the full model as the optimal model. Finally, 6 stations in the area are randomly selected, and the full model is used to interpolate the temperature and compared with ordinary kriging (OK) and inverse distance weighting (IDW).
2. Establishment of the model

2.1 Definition of Copula and C-vine
Sklar proposed Copula theory \([10]\), which pointed out that the composite distribution \(F\) with a univariate marginal distribution \(F_1, \ldots, F_n\). Then there exists a Copula \(C\) : \([0,1]^n \rightarrow [0,1]\) which holds that
\[
F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \tag{1}
\]
If \(F_1, \ldots, F_n\) in equation (1) is continuous, then \(C\) is unique.

In the framework of Copula, we often use C-vine and D-vine to describe the correlation of high-dimensional variables. According to its structural characteristics, this article is based on C-vine research.

Definition 1 (C-vine) \([11]\): \(V\) is called a C-vine structure built on \(n\)-dimensional variables, When the following three conditions are met

1) \(V = (T_1, \ldots, T_{n-1})\), \(T_i\) is a tree that constitutes a C-vine structure. We respectively denoted its node set and edge set as \(N_i\) and \(E_i\), \(i = 1, \ldots, n-1\)

2) The node set on \(T_1\) is \(N_1\) and the edge set is \(E_1\),

3) \(T_i (i = 1, \ldots, n-1)\) has only one root node connected to the \(n - i\) edges.

2.2 Definition of Spatial local C-vine
The definition of five-dimensional Spatial local C-Vine is as follows: Each station \(u_s (s = 1, \ldots, d)\) is considered a five-dimensional C-vine structure \(V^S = (T_1^S, T_2^S, T_3^S, T_4^S)\). The first root variable of C-vine at each station is denoted as \(s\). The variable that is closest to \(s\) is recorded as the second root variable as \(p_s\). The variable closer to \(s\) is recorded as the third root variable as \(q_s\). To record the third variable close to \(s\) as the fourth root variable as \(r_s\), and the fourth variable close to \(s\) as the fifth variable as \(t_s\). By definition, a five-dimensional Spatial local C-vine structure diagram can be obtained.

![Figure 1. Five-dimensional Spatial local C-vine structure map.](image)

2.3 Modeling of Spatial local C-vine
Before building the model, the latitude, longitude, and altitude information of each station needs to be obtained. Through latitude longitude and height, the horizontal and vertical distance between stations can be obtained. In order to reduce the impact of extreme values and heteroscedasticity, we use logarithmic transformation to get:
\[
D_1 = \ln(d_{i,0s}) \quad D_{21} = \ln(d_{j,0s}) \quad D_{22} = \ln(d_{j,i}) \quad D_{23} = \ln(d_{o2,i}) \quad D_{31} = \ln(d_{k,0s})
\]
\[
D_{32} = \ln(d_{k,i}) \quad D_{33} = \ln(d_{k,j}) \quad D_{34} = \ln(d_{o3,i}) \quad D_{35} = \ln(d_{o3,j}) \quad D_{41} = \ln(d_{m,0s})
\]
\[
D_{42} = \ln(d_{m,i}) \quad D_{43} = \ln(d_{m,j}) \quad D_{44} = \ln(d_{m,k}) \quad D_{45} = \ln(d_{o4,i}) \quad D_{46} = \ln(d_{o4,j})
\]
\[
D_{47} = \ln(d_{o4,k})
\]
Similarly, using logarithmic transformation to get:
\[
E_1 = \ln(e_{i,1s}) \quad E_{21} = \ln(e_{j,2s}) \cdots \quad E_{47} = \ln(d_{o4,k})
\]

Where \(i = s, j = p_s, k = q_s, m = r_s, o_s = p_s, q_s, r_s, t_s, o_2 = q_s, r_s, t_s, o_3 = r_s, t_s, o_4 = t_s\) \((s =\)
1, \cdots, d). \, d_{i,o}$ represents the sum of the distances from $s$ to $p_{s}, q_{s}, r_{s}, t_{s}$.

Model 1: We use the latitude, longitude, and altitude information of all stations to build a full model, $h_{l}^{\text{full}}, l = (1,2,3,4)$. The horizontal and vertical distance information of all stations are considered in the model.

\[
\begin{align*}
    h_{l}^{\text{full}}(i, o_{1}) &= \beta^{\text{full}}_{l,1} + \beta^{\text{full}}_{l,2} h_{l} + \beta^{\text{full}}_{l,3} v_{l} \\
    h_{l}^{\text{full}}(i, j, o_{2}) &= \beta^{\text{full}}_{l,4} + \sum_{i=1}^{3} \beta^{\text{full}}_{l,5} h_{2i} + \sum_{i=1}^{3} \beta^{\text{full}}_{l,6} v_{2i} \\
    h_{l}^{\text{full}}(i, j, k, o_{3}) &= \beta^{\text{full}}_{l,7} + \sum_{i=1}^{7} \beta^{\text{full}}_{l,8} h_{3i} + \sum_{i=1}^{7} \beta^{\text{full}}_{l,9} v_{3i} \\
    h_{l}^{\text{full}}(i, j, k, m, o_{4}) &= \beta^{\text{full}}_{l,10} + \sum_{i=1}^{7} \beta^{\text{full}}_{l,11} h_{4i} + \sum_{i=1}^{7} \beta^{\text{full}}_{l,12} v_{4i}
\end{align*}
\]

Model 2: We use the latitude and longitude information of all stations to build a horizontal distance model, $h_{l}^{\text{hor}}, l = (1,2,3,4)$. The horizontal distance information of all stations is considered in the model.

\[
\begin{align*}
    h_{l}^{\text{hor}}(i, o_{1}) &= \beta^{\text{hor}}_{l,1} + \beta^{\text{hor}}_{l,2} h_{l} \\
    h_{l}^{\text{hor}}(i, j, o_{2}) &= \beta^{\text{hor}}_{l,3} + \sum_{i=1}^{3} \beta^{\text{hor}}_{l,4} h_{2i} \\
    h_{l}^{\text{hor}}(i, j, k, o_{3}) &= \beta^{\text{hor}}_{l,5} + \sum_{i=1}^{7} \beta^{\text{hor}}_{l,6} h_{3i} \\
    h_{l}^{\text{hor}}(i, j, k, m, o_{4}) &= \beta^{\text{hor}}_{l,7} + \sum_{i=1}^{7} \beta^{\text{hor}}_{l,8} h_{4i}
\end{align*}
\]

Model 3: We use the height information of all stations to build vertical distance model, $h_{l}^{\text{vert}}, l = (1,2,3,4)$. The vertical distance information of all stations is considered in this model.

\[
\begin{align*}
    h_{l}^{\text{vert}}(i, o_{1}) &= \beta^{\text{vert}}_{l,1} + \beta^{\text{vert}}_{l,2} v_{l} \\
    h_{l}^{\text{vert}}(i, j, o_{2}) &= \beta^{\text{vert}}_{l,3} + \sum_{i=1}^{3} \beta^{\text{vert}}_{l,4} v_{2i} \\
    h_{l}^{\text{vert}}(i, j, k, o_{3}) &= \beta^{\text{vert}}_{l,5} + \sum_{i=1}^{7} \beta^{\text{vert}}_{l,6} v_{3i} \\
    h_{l}^{\text{vert}}(i, j, k, m, o_{4}) &= \beta^{\text{vert}}_{l,7} + \sum_{i=1}^{7} \beta^{\text{vert}}_{l,8} v_{4i}
\end{align*}
\]

2.4 Spatial local C-vine model selection

After constructing the above three models based on the latitude, longitude, and altitude information, the following steps will be used to select three models.

1. Construction of the Vine: The horizontal and vertical distance between each station are obtained and then the structure of Spatial local C-vine is determined according to the minimum value of $\sum_{l=1}^{4} h_{l}^{\text{full}}, \sum_{l=1}^{4} h_{l}^{\text{hor}}, \sum_{l=1}^{4} h_{l}^{\text{vert}}$.

2. The choice of Copula: We convert the distribution function of temperature data $y_{t}^{s}(t = 1 \cdots N, s = 1 \cdots d)$ of each station into Copula data $u_{t}^{s}$. We choose from some commonly used Copula function families, and only consider Copula function families with one or two parameters, which are respectively defined as $\theta$ and $\nu$.

3. Selection of model: After determining the structure and corresponding Copula of three Spatial local C-vine, the likelihood function of the temperature data of each station is as follows:

\[
\mathcal{L}_{s}(\theta, \nu \mid u_{t}^{s}, h_{t}^{p}, u_{t}^{q}, r_{t}^{s}, u_{t}^{r} z_{t}^{s}) = \epsilon_{s}(u_{t}^{p}, u_{t}^{q}, r_{t}^{s}, u_{t}^{r} z_{t}^{s} \mid \theta, \nu)
\]

After obtaining the respective likelihood functions, the composite likelihood function is constructed next. The weight of each station will be considered and recorded as: $\omega_{s}$, $s = 1, \cdots, d$. We defined $n_{s}$ as the number of $s$ in the first tree of all Spatial local C-vine, and $1/n_{s}$ as the reciprocal value of $n_{s}$. Then, the composite likelihood function of Spatial local C-vine can be obtained as:
\[ L_c(\theta^*, v^*|u^1, \ldots, u^d) = \prod_{s=1}^{N} \prod_{t=1}^{d} [L_s(\theta_s, v_s|u^t_s, u^{\hat{t}}_s)]^{w_s} \quad (6) \]

Where \( \theta^* = \{\theta|\theta \in \Theta, s = 1, \ldots, d\} \) and \( v^* = \{v|v \in v_s, s = 1, \ldots, d\} \).

Through the composite likelihood function, the optimal Spatial local C-vine model can be obtained.

2.5 Temperature interpolation based on Spatial local C-vine

After obtaining the optimal model, we also need to consider an observation station \( s \), and use the following steps to interpolate the temperature of observation station \( s \).

(1) According to the structure of Spatial local C-vine, observation stations \( p_s, q_s, r_s, t_s \) related to the observation station \( s \) can be obtained. Then we converted the air temperature data of these 4 observation stations into Copula data, and established following expression:

\[ \hat{\theta}_{ij|s}(i, j, o_s|\beta_{mod,3}) ; C_{ij|s} \]

(7)

Where, the meaning of \( i, j, k, m, o_1, o_2, o_3, o_4 \) has been described before. \( h^{mod}_l (l = 1, 2, 3, 4) \) is the optimal Spatial local C-vine model. \( \beta_{mod,l}(l = 1, 2, 3, 4) \) is the corresponding parameter. \( F^{-1}_l(\cdot) \) is the inverse function of the corresponding distribution function. \( C(\cdot) \) is the specified Copula. Similarly, the corresponding parameters \( \hat{v} \) can be obtained.

(2) In order to predict the temperature data of the observation station \( s \), we use \( p_s \) as the root variable of the first tree of Spatial local C-vine, then the corresponding parameters can be obtained by equation (8) as follows:

\[ \hat{\theta}_{ps|s}(p_s, q_s|\beta_{mod,1}) ; C_{ps|s} \]

(8)

Similarly, the corresponding parameters \( \hat{\theta}_{ps|p_s}, \hat{\theta}_{q_s|p_s}, \hat{\theta}_{r_s|p_s}, \hat{\theta}_{t_s|p_s} \) can be obtained. Then, through iterative calculation, the conditional distribution function \( F_{s|p_s} \) can be obtained:

\[ F_{s|p_s} = \frac{\partial C_{ps|p_s} \cdot F_{r_s|p_s} \cdot F_{t_s|p_s}}{\partial F_{s|p_s}} \]

(9)

(3) We conduct 1000 conditional inverse sampling, take the average value, and convert it into raw temperature data.

Therefore, for any observing station \( s \), the corresponding observing station \( p_s, q_s, r_s, t_s \) can be obtained through Spatial local C-vine. According to the temperature data of these four observation stations, the temperature at station \( s \) can be predicted.

3. Empirical analysis

3.1 Spatial local C-vine structure of the optimal model

Daily average temperature of 73 stations in Yunnan Province from 1980 to 2010, as well as the latitude, longitude and altitude information of each station were obtained from the China Meteorological Data Network(https://data.cma.cn/) in this article. Then, we take the logarithm of the horizontal distance and vertical distance between each station, and construct a full model, a horizontal distance model and a vertical distance model. Equation (6) is used to calculate the composite likelihood function, and then we get maximum likelihood value, AIC and BIC of each model in the following table.
Table 1. Values of maximum likelihood, AIC and BIC of three Spatial local C-vine.

| Model                      | Method   | $h_t^{\text{full}}$ | $h_t^{\text{hori}}$ | $h_t^{\text{vert}}$ |
|----------------------------|----------|----------------------|----------------------|---------------------|
| Maximum likelihood value   |          | 86595                | 74820                | 73471               |
| AIC                        |          | -172994              | -149626              | -134696             |
| BIC                        |          | -172611              | -149597              | -134310             |

From the perspective of likelihood, we can know that the likelihood value of the full model is the largest in the above three models. From the perspective of AIC and BIC, the value of the full model is the smallest in Table 1.

According to Spatial local C-vine structure of full model, the C-vine structure between 73 meteorological stations in Yunnan Province can be obtained, as shown in the following figure.

Figure 2. The Spatial local C-vine structure of full model.

It can be seen from the figure that the connection is close between the adjacent stations, indicating that the temperature data between the adjacent stations is similar. The number of each station indicates that the first tree of all Spatial local C-vine contains the number of the station, that is, $n_s$.

After Spatial local C-vine structure is obtained, the distribution function of the temperature data at each station is converted into Copula data. Then AIC and BIC are used to select the Pair-Copula on each edge. Pair-Copula of each edge between some stations is given, as shown in table 2:

Table 2. Pair-Copula (PC) corresponding to each edge of 6 stations.

| PC stations | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-------------|----|----|----|----|----|----|----|----|----|----|
| 1           | C  | C  | C  | F  | T  | F  | T  | C  | T  | T  |
| 2           | C  | C  | C  | G  | C  | T  | T  | T  | T  | T  |
| 3           | C  | C  | C  | F  | T  | T  | T  | T  | T  | T  |
| 4           | C  | F  | C  | F  | T  | T  | F  | T  | G  |    |
| 5           | C  | C  | C  | F  | T  | T  | T  | T  | T  |    |
| 6           | F  | C  | C  | F  | T  | T  | T  | T  | T  | T  |
Note: (C is Clayton Copula; F is Frank Copula; T is t-Copula.)
We designate Clayton Copula, Frank Copula, t-Copula represents the first three edges of the first tree, the fourth edge of the first tree and each edge of the following three trees.

3.2 Spatial interpolation and precision comparison
In order to verify the practicability of full model, a test set and a verification set are selected from 73 national weather stations in Yunnan Province. 6 stations are randomly sampled as validation sets. Through Spatial local C-vine of the full model, four stations corresponding to six stations can be obtained. We use equation (2) to calculate the corresponding $h^\text{full}_l$, $(l = 1, \ldots, 4)$, and then bring this result and the result of Table 2 into equation (8) to calculate the corresponding parameters. The corresponding distribution function is calculated by equation (9). Finally, convert them into temperature data. The following figure shows temperature prediction map of the 6 stations:

![Figure 3. Comparison of predicted and actual values at 6 stations.](image)

It can be drawn from Figure 3 that the actual temperature curves of the six stations are close to the predicted temperature curves, indicating that the interpolation accuracy is good. We have observed the two temperature curves of Yuxi almost overlap, indicating that the interpolation prediction effect is the best. This is because Yuxi is located in the central part of Yunnan Province, and there are many weather stations around it.

Next, we calculated the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of these 6 stations, and compared it with the Ordinary Kriging method (OK) and Inverse Distance Weighting method (IDW). The results are shown in Table 3.

| Method       | full model | OK          | IDW        |
|--------------|------------|-------------|------------|
| RMSE         | 0.26520    | 0.33489     | 0.36646    |
| MAE          | 0.01685    | 0.02097     | 0.02289    |

It can be drawn from Table 3 that the full model interpolation method based on Spatial local C-vine has a smaller MAE and RMSE than OK and IDW. The verification results show that the full model based on Spatial local C-vine has more prediction accuracy than the other two methods.
4. Conclusion
Adopting full model of Spatial local C-vine to interpolate the temperature in Yunnan Province, we can obtain the following conclusions: (1) The predicted temperature is closest to the actual temperature of Yuxi. (2) The temperature prediction curve and actual curve of the remaining 5 stations almost coincide.

By calculating MAE and RMSE of these 6 stations, and then compared with OK and IDW, the full model has better prediction accuracy. This shows that the full model based on Spatial local C-vine has a strong applicability in the study area, and provides a theoretical basis for temperature interpolation.

References
[1] Aalto J, Pirinen P, Heikkinen J, et al. (2013) Spatial interpolation of monthly climate data for Finland: comparing the performance of kriging and generalized additive models. Theoretical and applied climatology, 112:99-111.
[2] Li K, Zhou M, Chen J, et al. (2019) An approach of improved gradient plus inverse distance squared for spatial interpolation of temperature. Journal of University of China Academy of Science, 36:491-497.
[3] Roznik M, Brock P, Porth L. (2019) Improving agricultural microinsurance by applying universal kriging and generalised additive models for interpolation of mean daily temperature. The Geneva Papers on risk and Insurance-Issues and practice, 44:446-480.
[4] Roger B, Nelsen. (2000). An Introduction to Copulas. Technometrics, 42(3):317.
[5] [11] Bedford T, Cooke R M. (2001) Probability density decomposition for conditionally dependent random variables modeled by Vines. Annals of Mathematics and Artificial Intelligence, 32:245-268.
[6] Li Y, Hou J, Huang C. (2017) Spatial Downscaling Land Surface Temperature based on Copula. Remote Sensing Technology and Application, 5:818-824.
[7] Leng M, Song W, Bai H, et al. (2019) Copula function-based design of daily temperature for sponge city. Water Resources and Hydropower Engineering, 50:1-12.
[8] Ye M, Wang R, Ding Y, et al. (2017) Interaction of temperature and rainfall in Anhui province based on copula function. Resources and environment in the Yangtze basin, 26:110-117.
[9] Erhardt T M, Czado C, Schepsmeier U. (2015) Spatial composite likelihood inference using local C-vine. Journal of Multivariate Analysis, 138:74-88.
[10] Sklar A. (1959) Fonctions de repartition an dimensions et leurs marges. Publication de l'Institutde de Statistique l'Universite de Paris, 8:229-231.