Network Harmonization of Spatial Interval-Valued Data

Emmanuel Sarfo Fosu [1] Soorok Ryu [2] GyuWon Lee [3] Joon Jin Song [1]

[1]Department of Statistical Science, Baylor University, Waco, Texas 76798-7140, USA
[2]Center for Atmospheric REmote Sensing (CARE), Department of Atmospheric Sciences, Kyungpook National University, Republic of Korea
[3]BK21 Weather Extremes Education & Research Team, Department of Atmospheric Sciences, Center for Atmospheric REmote sensing (CARE), Kyungpook National University, Republic of Korea

Introduction

Numerous spatial data sources emerge through spatial networks that operate with varying objectives and protocols leading to heterogeneity in data. In reconciling, researchers face hurdles in addressing these heterogeneities [2].

Data harmonization is merging data from diverse origins targeting the same variable while ensuring compatibility and consistency. This offers several advantages but remains relatively underexplored in the spatial world.

Symbolic data analysis is a concept pioneered by [1], and it can be obtained naturally or through processing. **Interval-valued data (IVD)** is a type of data where each observation is represented as an interval, capturing inherent uncertainty and variability. For instance, consider two intervals, $(22, 26)$ and $(14, 34)$, both centered at 24 but with different widths. Interval features are used in the analysis, typically represented as (min, max) or (center, range).

Motivation

The methodologies and applications of IVD are emerging in spatial statistics, as evidenced in [3]. The goal of this study is:

- To achieve integrated data by effectively incorporating information from different networks while preserving and maximizing information retention.
- To perform efficient and accurate spatial predictions in unsampled locations.

Harmonization Procedure

Algorithm : Procedure to harmonize and predict Spatial Interval-Valued Data (SIVD)
Data: Data observed from heterogeneous networks.
Result: Spatial prediction with harmonized data.
1 Identify the spatial overlapping networks.
2 Estimation of differences of the spatial network.
3 **for** *each differenced data* **do**
4 Test network heterogeneity
5 **if** *differences are significant* **then**
6 Bias estimation and generating debiased data
7 Perform kriging with the harmonized data.
8 **else**
9 Combine network data and perform kriging.

Methodology

The observed data, $\mathbf{Y}(\mathbf{s})_i^{CR}$ under SIVD harmonization model is defined:

$$\mathbf{Y}(\mathbf{s})_i^{CR} = \boldsymbol{\mu}(\mathbf{s})^{CR} + \mathbf{b}_i^{CR} + \mathbf{e}(\mathbf{s})^{CR} + \boldsymbol{\delta}(\mathbf{s})^{CR} \tag{1}$$

where, $\boldsymbol{\mu}(\mathbf{s})^{CR}$ is mean, $\mathbf{b}_i^{CR}$ is network bias, $\mathbf{e}(\mathbf{s})^{CR}$ is spatial error, $\boldsymbol{\delta}(\mathbf{s})^{CR}$ is random error and the ^{CR} denotes a vector of center and range.

The true or target process is then represented as:

$$\mathbf{Z}(\mathbf{s})_i^{CR} = \boldsymbol{\mu}(\mathbf{s})^{CR} + \mathbf{e}(\mathbf{s})^{CR} + \boldsymbol{\delta}(\mathbf{s})^{CR}, \tag{2}$$

We begin by using the original values from j to predict the location of networks i to obtain $\mathbf{q}_{ij}^{CR}(\mathbf{s}) = \mathbf{Y}_i^{CR}(\mathbf{s}) - \hat{\mathbf{Y}}_j^{CR}(\mathbf{s})$. Consequently, $E(\mathbf{q}_{ij}^{CR}(\mathbf{s})) = \mathbf{b}_i^{CR} - \mathbf{b}_j^{CR}$.

The weighted average estimate for the differences and variance for the features are given as:

$$\hat{q}_{w_{ij}}^{CR}(\mathbf{s}) = \frac{\mathbf{1}^T\mathbf{W}^{CR}\left(\mathbf{y}_i^{CR}(\mathbf{s}) - \hat{\mathbf{y}}_j^{CR}(\mathbf{s})\right)}{\mathbf{1}^T\mathbf{W}^{CR}\mathbf{1}}, \tag{3}$$

$$v_{q_{ij}^{CR}(\mathbf{s})} = \left[\frac{\mathbf{1}^T\mathbf{W}^{CR}\left(\mathbf{y}_i^{CR}(\mathbf{s}) - \hat{\mathbf{y}}_j^{CR}(\mathbf{s})\right)^2 - \mathbf{1}^T\mathbf{W}^{CR}(\bar{\mathbf{q}}_{y-\hat{y}}^{CR})^2}{(n_i - 1)\mathbf{1}^T\mathbf{W}^{CR}\mathbf{1}}\right]. \tag{4}$$

Methodology Cont'd

The linear model for the relationship between the expected difference and bias is given as

$$\mathbf{q}^{CR} = \mathbf{D}^{CR}\mathbf{b}^{CR} + \boldsymbol{\varepsilon}^{CR}, \tag{5}$$

where, $\mathbf{D}^{CR}$ is a relationship matrix. We estimate the optimal values of the $\mathbf{b}^{CR}$ by

$$\underset{\mathbf{b}_C,\mathbf{b}_R}{\arg\min}\left\|(\mathbf{W}^C)^{\frac{1}{2}}\cdot\boldsymbol{\varepsilon}^C\right\|^2 + \left\|(\mathbf{W}^R)^{\frac{1}{2}}\cdot\boldsymbol{\varepsilon}^R\right\|^2. \tag{6}$$

We solve the systems of equations from Eq. 6 to obtain the biases by considering one network as the reference. These estimates for both features are given as:

$$\hat{\mathbf{b}}^{CR} = \left(\mathbf{D}^{CR'}\boldsymbol{\Omega}_{CR}^{-1}\mathbf{D}^{CR}\right)^{-1}\mathbf{D}^{CR'}\boldsymbol{\Omega}_{CR}^{-1}\mathbf{q}^{CR} \qquad \mathbf{Var}(\hat{\mathbf{b}}^{CR}) = \boldsymbol{\sigma}^2_{\mathsf{MSE}_{\mathbf{q}^{CR}}}\left(\mathbf{D}^{CR'}\boldsymbol{\Omega}_{CR}^{-1}\mathbf{D}^{CR}\right)^{-1}, \tag{7}$$

where, $\boldsymbol{\Omega}_{CR} = \mathsf{Diag}\left(\mathbf{v}_{q^{CR}(\mathbf{s})}\right)$.

Once the network systematic biases are estimated, they are eliminated from the observed model to achieve the true underlying process.

$$\mathbf{Z}(\mathbf{s})_i^{CR} = \mathbf{Y}(\mathbf{s})_i^{CR} - \mathbf{b}_i^{CR} \implies \mathbf{Z}(\mathbf{s})^{CR} = \begin{pmatrix}\mathbf{vec}\left([\mathbf{Z}(\mathbf{s})_1^C\ \mathbf{Z}(\mathbf{s})_2^C\ \ldots\ \mathbf{Z}(\mathbf{s})_k^C]^T\right)\\ \mathbf{vec}\left([\mathbf{Z}(\mathbf{s})_i^R\ \mathbf{Z}(\mathbf{s})_2^R\ \ldots\ \mathbf{Z}(\mathbf{s})_k^R]^T\right)\end{pmatrix} = \begin{pmatrix}\mathbf{Z}^C\\ \mathbf{Z}^R\end{pmatrix} \tag{8}$$

After data harmonization, the resulting interval prediction at location $\mathbf{s}_0$ is

$$\hat{\mathbf{Z}}(\mathbf{s_0})^{CR} = \left(\mathbf{w}(s_0)_C^T\mathbf{Z}^C - \frac{1}{2}\exp\left\{\mathbf{w}(s_0)_R^T\mathbf{Z}^R\right\},\quad \mathbf{w}(s_0)_C^T\mathbf{Z}^C + \frac{1}{2}\exp\left\{\mathbf{w}(s_0)_R^T\mathbf{Z}^R\right\}\right). \tag{9}$$

We use the IVD RMSPE proposed by [3] as a means to assess predictions for both bounds and internal variations.

Simulation Study

- **Setup**
  - Simulated data for 1000 locations for the interval features from a Gaussian process.
  - Network 1 is set as the reference network.
  - Four networks with varied no. of observations for reference (high, equal, and low).
  - We considered the following bias for the four networks: $b_c = (0, 2, 4, 6)$, $b_{lr} = (\log(0.001), \log(1), \log(2), \log(3))$
  - Range $\boldsymbol{\phi_C} = (15, 20, 25, 30)$ and $\boldsymbol{\phi_R} = (2, 4, 6, 8)$; Partial sill $\boldsymbol{\sigma_C^2} = (3, 5, 7, 9)$ and $\boldsymbol{\sigma_R^2} = (0.3, 0.5, 0.8, 1.0)$
  - Assessed kriging with Harmonized, Pooled, and Reference data.
- **Results**
  - Systematic bias estimation is fairly robust to varied spatial parameters.
  - There is higher variability in the estimation when the reference network holds the smallest proportion of data.
  - The precision of the bias estimates is highest when the reference network holds the largest proportion of the data being analyzed.
  - As the spatial parameters increase, the model prediction accuracy decreases.
  - The model prediction is inversely proportional to the size of the reference network.

Table 1. Lower (L), Upper (U), Combined (C), and Range (R) RMPSE for data generated with ϕ_C = 20, ϕ_R = 4, σ_C^2 = 5, σ_R^2 = 0.5.

	$\mathbf{RMSPE_L}$			$\mathbf{RMSPE_U}$			$\mathbf{RMSPE_C}$			$\mathbf{RMSPE_R}$		
Prop	Ham	Net1	Pool	Ham	Net1	Pool	Ham	Net1	Pool	Ham	Net1	Pool
high	**0.199**	0.662	2.734	**0.198**	0.662	2.369	**0.280**	0.936	3.622	**0.357**	1.308	1.235
equal	**0.200**	0.665	3.833	**0.200**	0.664	4.194	**0.282**	0.940	5.685	**0.361**	1.306	1.392
low	**0.201**	0.682	4.812	**0.201**	0.681	6.115	**0.284**	0.964	7.804	**0.362**	1.324	1.884

Real data Application

- We illustrate the harmonized kriging on Spatial interval-valued temperature data gathered from three distinct monitoring networks in South Korea.
- Networks consist of the Korea Meteorological Administration (KMA), the Korea Forest Service (KFS), and the Rural Development Administration (RDA).
- Data comprises $1,147$ observations recorded on July 22, 2020.
- KMA is used as the reference network.

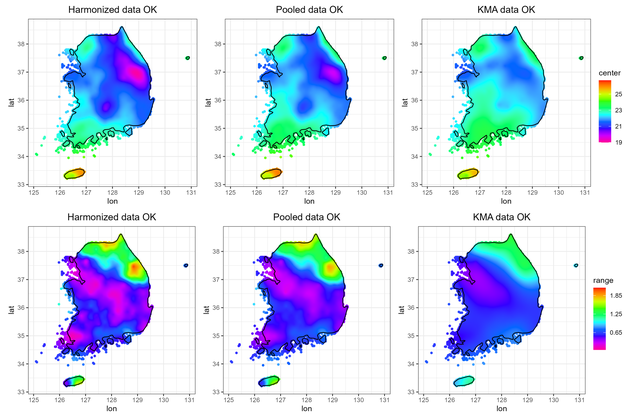

Figure 1. Center (top) and Range (bottom) temperature prediction at unsampled locations in South Korea.

Table 2. Lower(L), Upper(U), Combined(C), and Range (R) RMSPE for kriging with harmonized, pooled and KMA SIVD temperature.

	Harmonized	KMA	Pooled
$RMSPE_L$	$\mathbf{1.0448}$	1.1122	1.1461
$RMSPE_U$	$\mathbf{1.0573}$	1.1491	1.1929
$RMSPE_C$	$\mathbf{1.4864}$	1.5992	1.6542
$RMSPE_R$	$\mathbf{0.4264}$	0.4453	0.4409

Discussion and Conclusion

- We proposed a statistical network harmonization for spatial interval-valued data.
- Simulation and a real data application demonstrated superior predictive performance using harmonized data.
- Further research could include:
  - extending SIVD harmonization into the non-stationary framework.
  - developing a multivariate approach to data harmonization.
  - considering a downscaler approach to harmonization.
  - obtaining advanced mode of data visualization.

Acknowledgments

This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMI2022-00310.

References

[1] E. Diday. Introduction à l'approche symbolique en analyse des données. premières journées symbolique-numerique. In *Workshop. CEREMADE Laboratory*, 1987.

[2] J. O. Skøien, O. P. Baume, E. J. Pebesma, and G. B. M. Heuvelink. Identifying and removing heterogeneities between monitoring networks. *Environmetrics*, 21(1):66–84, 2010.

[3] A. Workman and J. J. Song. Spatial analysis for interval-valued data. *Journal of Applied Statistics*, 0(0):1–15, 2023.

emmanuel_sarfofosu1@baylor.edu