Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Data Scarcity or low Representativeness?: What hinders accuracy and precision of spatial interpolation of climate data?

Avit Kumar Bhowmik
Institute for Environmental Sciences, University of Koblenz-Landau
Fortstraße 7
76829 Landau in der Pfalz, Germany
bhowmik@uni-landau.de

Ana Cristina Costa
ISEGI, Universidade Nova de Lisboa
1070-312 Lisbon, Portugal
ccosta@isegi.unl.pt

Abstract

Data scarcity is a major scientific challenge for accuracy and precision of spatial interpolation of climatic fields, especially in climate-stressed developing countries. Methodologies have been suggested for coping up with data scarcity but data have rarely been checked for their representativeness of corresponding climatic fields. Here, influences of number and representativeness of climate data on accuracy and precision of their spatial interpolation were investigated and compared. Two precipitation and temperature indices were computed for a long time series in Bangladesh, which is a data scarce region. The representativeness was quantified by dispersion in the data and the accuracy and precision of spatial interpolation were computed by four commonly used error statistics derived through cross-validation. The precipitation data showed very little and sometimes null representativeness whereas the temperature data showed very high representativeness of the corresponding fields. Consequently, interpolated precipitation surfaces showed little accuracy and precision whereas temperature surfaces showed high accuracy and precision despite the scarce data. The results indicate that representativeness of climate data, i.e. variability of climate phenomenon, is more crucial than the number of data for accuracy and precision of spatial interpolation and should be treated with higher importance.
Keywords: Precipitation, temperature, point density, spatial interpolation, error statistics, regression.

1 Introduction

Spatial interpolation is an essential tool for continuously deriving climate information over space based on data at particular locations. Low accuracy and precision in spatial interpolation occurs in regions with a few climate data, e.g. in developing regions where the available number of data is often technologically and economically constrained [1, 2]. However, representativeness of corresponding climatic fields is also one of the important data characteristics and may ensure satisfactory accuracy and precision in spatial interpolation in data scare regions [3].

Consequently, the research question of this study was - can high representativeness of climate data ensure satisfactory accuracy and precision in spatial interpolation despite their scarcity?

2 Study area

In Bangladesh, only 34 meteorological stations currently report daily precipitation and temperature over 147,570 km^2 areal extent [4], and thus distinguish it as a data scarce region (Figure 1 (a)). During the period of 1948-2007, there is a gradual increase in the number of data locations for precipitation and temperature, i.e. from 8 to 32 and from 10 to 34, respectively (Figure 1(b)).

3 Materials and Methods

The daily precipitation and temperature data during 1948-2007 in Bangladesh were used. Two annual climate indices – the annual total precipitation in wet days (PRCPTOT) and the yearly maximum value of the daily maximum temperatures (TXx) were computed [5, 6].

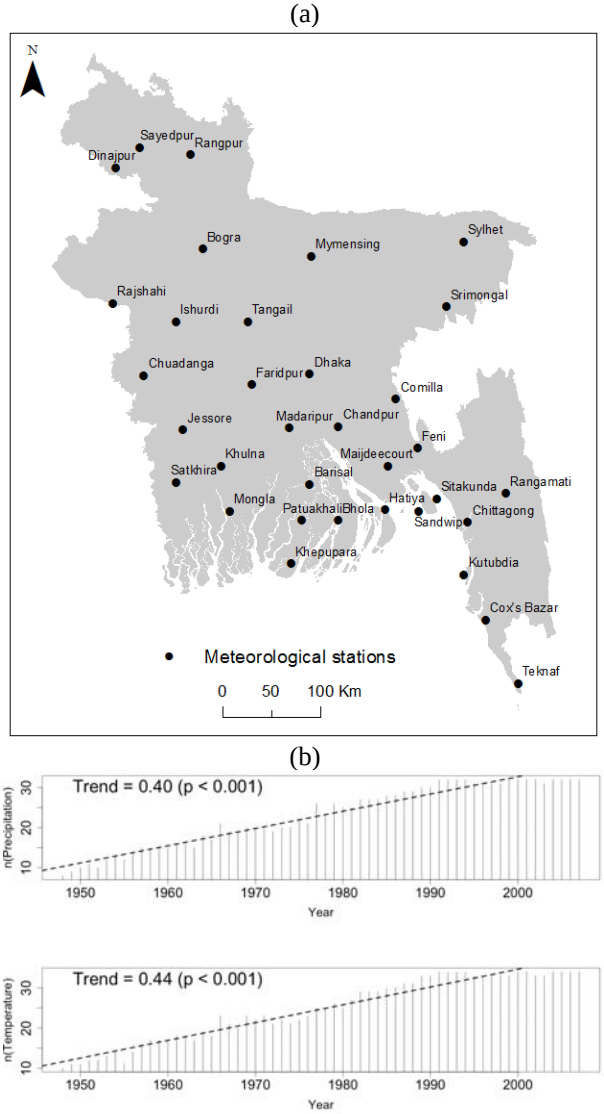
The representativeness of the climate indices was quantified by the measure of the regional coefficient of variation (k, expressed in terms of a percentage) [7].

Required number of data (n) for obtaining satisfactorily accurate and precise interpolated surfaces, i.e. root mean square error (RMSE) of the interpolated values lower than or equal to 5% of the regional mean (M) of the observed indices [8, 9], was estimated and compared to the available n according to [10, 11].

The Universal Kriging (UK) spatial interpolation model [12] was applied using the 'gstat' package [13]

in R [14]. The pooled variogram [15] parameters for three successive periods: 1948-1972, 1973-1992 and 1993-2007, taken from Bhowmik [16], were fitted to the UK model. The accuracy and precision of spatial interpolation were measured by four error statistics: (1) root mean square error (RMSE), 2) mean absolute error (MAE), 3) systematic root mean square error (RMSEs), and 4) unsystematic root mean square error (RMSEu). These were computed using the 'gstat' [13] and 'hydroGOF' [17] packages.

Figure 1: (a) Spatial distribution of the 34 current meteorological stations in Bangladesh and (b) increasing number of the data for precipitation (n(Precipitation)) (top) and temperature (n(Temperature)) (bottom) during 1948-2007.



Source: Author produced from the data [11].

The bivariate normality between the available n and the k was tested using the Henze-Zirkler's Multivariate Normality Test [18] in the 'MVN' package of R [19] and they were significantly ($p<0.001$) well fitted. The explained variability in the available n (generalized linear model with the 'poisson' family) and k (simple linear regression model) by each other were analyzed and the residuals of both models were extracted to obliterate the effects of the available n and the k on each other.

Finally, the bivariate normality in the residuals of n and k separately paired with each of the error statistics were tested [18]. Consequently, four simple linear regression models were fitted with the residuals of n and k as predictors (independent variables), separately, to predict the four spatial interpolation error statistics for both the indices (PRCPTOT and TXx). The percentage of variability in the response variables explained by the residuals of n and k was evaluated by the adjusted coefficient of determination (R^2) and through the statistical significance ($p<0.05$) of their corresponding regression parameters.

4 Results and Discussion

An average k of 41% was observed for PRCPTOT resulting from a range of 24.57-59.51% whereas for TXx the average k was 6.2% with a range of 3.26-23.97% (Table 1). Thus, the TXx data were mostly representative of the climatic field, whereas PRCPTOT data were unrepresentative.

For PRCPTOT, the number of available data did not meet the requirement for obtaining satisfactory accuracy and precision of spatial interpolation according to the computed k in any of the time steps (Table 1). On the contrary, in 43 time steps out of 60 (72%), the available n of TXx data met the requirement for satisfactory accuracy and precision.

The available n and k could significantly explain each other though the explained variability and slopes were very low (Table 2). On an average, k explains much higher variability of the error statistics than the available n and showed statistical significance when n and k were independent of each other (Table 3). The k explains 39.67% [78.16%] of the variability in the RMSE of spatial interpolation of PRCPTOT [TXx] and the corresponding regression parameters were statistically significant. Complementarily, n explains only 13.56% [3.38%] of the variability in the RMSE of PRCPTOT [TXx], thus its regression parameters are not statistically significant. More than 70% of the variability in each of the error statistics of TXx is explained by k except for RMSEu. Thus, the representativeness of the data, i.e. the variability of the

climate phenomenon, is more crucial than the number of data for ensuring accuracy and precision. The number of data is weakly related with the accuracy and precision, whereas their representativeness has a significant relation. The results are in line with [20].

Table 1: Computed representativeness (k – coefficient of variation) of the climate indices (PRCPTOT and TXx) at every time step (years) of the study period and corresponding available and required number of data (n) according to the measured k (Kelley, 2007; Lynch and Kim, 2010) for satisfactorily accurate and precise (RMSE - root mean square error <=5% of the regional mean of the indices (M)) spatial interpolation.

Years	k (%)		Available n		Required n (RMSE<=5%*M*)	
	PRCPTOT	TXx	PRCPTOT	TXx	PRCPTOT	TXx
1948	53.59*	3.69	8	10	985	10
1949	47.38	8.04	9	11	758	43
1950	54.97*	3.35	10	11	985	8
1951	53.79*	4.41	11	12	985	12
1952	37.94	6.44	10	12	423	35
1953	32.92	3.26	12	13	303	7
1954	36.71	4.38	13	14	423	11
1955	31.22	6.29	12	11	303	34
1956	24.57	6.85	14	14	137	37
1957	33.49	9.97	15	16	303	54
1958	46.61	7.52	15	17	758	41
1959	34.41	7.06	15	17	303	38
1960	42.15	5.68	15	17	573	31
1961	59.51*	4.90	16	18	985	16
1962	36.98	4.58	16	18	423	15
1963	36.05	4.67	15	17	423	15
1964	40.71	4.99	18	19	573	16
1965	42.57	7.49	17	18	573	40
1966	46.47	7.86	21	23	758	42
1967	46.17	23.97	19	20	758	137
1968	43.45	4.95	19	20	573	16
1969	44.49	8.49	20	23	573	46
1970	34.88	4.98	20	22	303	16
1971	53.34*	3.77	20	23	985	12
1972	45.54	4.98	19	21	758	16
1973	40.05	7.49	20	22	573	40
1974	43.25	5.19	20	21	573	21
1975	43.85	7.72	22	22	573	42
1976	42.05	8.38	21	23	573	45
1977	44.16	4.80	26	25	573	15
1978	39.50	4.86	23	25	423	16
1979	51.35*	5.02	26	26	985	26
1980	44.85	4.91	25	25	573	16
1981	44.74	3.67	25	27	573	10
1982	47.11	3.92	27	29	758	13
1983	42.48	4.60	27	29	573	15
1984	32.84	5.37	27	29	303	29
1985	39.56	6.76	28	30	423	30
1986	26.15	5.75	28	30	209	30
1987	36.17	5.90	29	31	423	31
1988	37.99	5.67	29	31	423	31
1989	47.31	7.24	30	33	758	32
1990	39.06	4.53	30	33	423	14
1991	37.57	4.09	32	34	423	13
1992	39.57	5.94	32	34	423	32
1993	36.75	3.97	32	34	423	13
1994	53.90*	5.64	32	34	985	30
1995	32.16	5.95	31	33	303	32
1996	38.33	5.06	31	33	423	27
1997	32.61	4.07	31	33	303	13

1998	42.04	4.35	31	33	573	14
1999	34.17	4.67	32	33	303	15
2000	49.42	3.12	32	34	758	8
2001	51.07*	3.16	32	34	985	8
2002	29.77	5.39	32	33	209	29
2003	48.09	12.81	31	33	758	69
2004	28.25	10.95	32	34	209	59
2005	37.85	5.18	32	34	423	28
2006	40.27	3.53	32	34	573	8
2007	28.55	4.15	32	34	209	13

*Null representativeness, i.e. k>50%

Source: Author produced.

Table 2: Coefficients of the simple linear regression model and generalized linear models with poisson family fitted to the representativeness (k – coefficient of variation) and the available number of data (n) of the climate indices (PRCPTOT and TXx), respectively, where were the n and k were respectively the predictors. The intercept, slope and the adjusted explained variability of the models are presented with the coefficients' statistical significance ($p<0.05$).

Response variables	Predictor variables							
	n							
	PRCPTOT				TXx			
	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
k (Simple linear regression model)	45.38*	-0.18*	0.13	1.91	7.54*	-0.05*	0.05	0.31
	k							
	PRCPTOT				TXx			
	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
n (Generalized linear model with the poisson family)	3.47*	-0008*	0.004	4.44	3.30*	-0.02*	0.009	1.98

*Statistically significant at $p<0.05$

Source: Author produced.

Table 2: Coefficients of the simple linear regression models fitted to the error statistics (RMSE – root mean square error, MAE – mean absolute error, RMSEs – systematic root mean square error and RMSEu – unsystematic root mean square error), where the representativeness (k –coefficient of variation) and the available number of data (n) of the climate indices (PRCPTOT and TXx) were separately the predictors. The intercept, slope and the adjusted explained variability of the linear regression models are presented with the coefficients' statistical significance ($p<0.05$).

Response variables	Predictor variables							
	Residuals of k							
	PRCPTOT				TXx			
Error Statistics	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
MAE	102.36*	7.91*	1.64	27.53	-0.09*	0.23*	0.02	76.29
RMSE	86.42*	11.59*	1.84	39.67	-1.09*	0.49*	0.03	78.16
RMSEs	-18.85*	9.36*	3.14	11.82	-2.21*	0.56*	0.05	72.34
RMSEu	88.44*	8.24*	2.84	11.14	-0.42	0.32*	0.03	60.88
	Residuals of n							
	PRCPTOT				TXx			
Error Statistics	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
MAE	579.18	-6.60	1.72	17.92	2.29	-0.04	0.01	3.22
RMSE	723.43	-7.01	2.19	13.56	3.08	-0.05	0.03	3.38
RMSEs	539.07	-7.56	3.20	7.18	2.57	-0.05	0.03	2.28
RMSEu	501.66	-3.27	2.99	0.30	2.28	-0.03	0.02	2.14

*Statistically significant at $p<0.05$

Source: Author produced.

It can be argued that the number of observations is somehow affecting the accuracy and precision of spatial interpolation of the indices, despite not being significant in general, and that its influence is considerably lower than the representativeness (Table 2) [21, 22]. Hence, in regions with abundant data, satisfactory accuracy and precision could be obtained without taking their representativeness into account [23]. However, in data scarce regions, the representativeness of the climate data should be treated with high importance.

References

[1] P. Dumolard. Uncertainty from spatial sampling: A case study in the French Alps. In H. Dobesch, P. Dumolard and I. Dyras, editors, Spatial Interpolation for Climate Data. The Use of GIS in Climatology and Meteorology, pages 57—70. ISTE Ltd., London, 2007.

[2] P. D. Wagner, P. Fiener, F. Wilken, S. Kumar, K. Schneider. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. J. Hydro., 464-465:388-400, 2012.

[3] T. Hill and P. Lewicki. Statistics: Methods and Applications: a Comprehensive Reference for Science, Industry, and Data Mining. In StatSoft. Tulsa, OK, 2006.

[4] DMICCDMP - Disaster Management Information Center of Comprehensive Disaster Management Program. Bangladesh Meteorological Department. http://www.bmd.gov.bd/index.php. Accessed 1 May 2013.

[5] P. Frich, L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, T. Peterson. Observed coherent changes in climatic extremes during the second half of the twentieth century. Clim. Res., 19:193–212, 2002.

[6] T. C. Peterson, C. Folland, G. Gruza, W. Hogg, A. Mokssit, N. Plummer. Report WCDMP-47, WMO-TD 1071. In Report on the activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998–2001. World Meteorological Organization, Geneva, 2001.

[7] M. G. Vangela. Confidence intervals for a normal coefficient of variation. Am. Stat. 50(1):21-26, 1996.

[8] J. J. Carrera-Hernández and S. J. Gaskin. Spatiotemporal analysis of daily precipitation and temperature in the Basin of Mexico. J. Hydro. 336:231-249, 2007. DOI: 10.1016/j.jhydrol.2006.12.021.

[9] P. C. Kyriakidis, J. Kim and N. L. Miller. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. J. Clim. 40(11):1855-1877, 2001.

[10] K. Kelley. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. Behav. Res. Methods. 39(4):755-766, 2007.

[11] R. M. Lynch and B. Kim. Sample size, the margin of error and the coefficient of variation. InterStat 4, 2010.

[12] R. Kerry and M. A. Oliver. Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Prec. Agri. 9(1-2):33-56, 2008.

[13] E. Pebesma. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30:683-691, 2004.

[14] R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna. http://www.R-project.org, Accessed 1 March 2014.

[15] R. S. Bivand, E. J. Pebesma and V. Gómez-Rubio. Applied Spatial Data Analysis with R. Springer Science and Business Media, LLC: New York, 2008.

[16] A. K. Bhowmik. A comparison of Bangladesh climate surfaces from the geostatistical point of view. *ISRN. Meteorol.*, 2012:353408, 2012.

[17] M. Zambrano-Bigiarini. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series, r package version 0.3-7. http://cran.r-project.org/web/packages/hydroGOF/, 2011.

[18] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. Commun. Stat. Theory. Methods. 19(10):3595-3617, 1990.

[19] S. Korkmaz. MVN: Multivariate Normality Tests. http://cran.r-project.org/web/packages/MVN/index.html, 2013.

[20] B. C. Hewitson and R. G. Crane. Gridded area-averaged daily precipitation via conditional interpolation. J. Clim. 18:41-57, 2005.

[21] A. Basistha, N. K. Goel, D. S. Arya and S. K. Gangwar. Spatial pattern of trends in Indian sub-divisional rainfall. Jalv. Sam., 22:47-57, 2007.

[22] M. Radziejewski and Z. W. Kundzewicz. Detectability of changes in hydrological records. Hydrol. Sci. J., 49: 39-51, 2004.

[23] U. Haberlandt. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. J. Hydro., 332:144-157, 2007.