УДК 528.942:551.577.21

Alexander Mkrtchian, PhD, docent
e-mail: alemkrt@gmail.com
Ivan Franko National University of Lviv

PROCCESSING AND GEOSTATISTICAL INTERPOLATION OF DATA ON ANNUAL PRECIPITATION FOR THE METEOSTATIONS OF WESTERN UKRAINE

There is a significant demand for accurate and reliable spatially distributed precipitation data. Modeling and mapping of precipitation fields are best achieved by geostatistical interpolation procedures that also consider explanatory variables like the terrain morphometric parameters that influence the precipitation distribution. The purpose of this research was to create an accurate regional precipitation map by interpolating data on average annual precipitation sums gained through summarizing records of 50 meteorological stations located in Western Ukraine. Daily data have been downloaded from open GHCN database, then preprocessed and summarized in R to obtain average annual precipitation sums for each station. Auxiliary data on terrain morphometric parameters have been gathered by preprocessing of SRTM Version 4.1 DEM. Data were then interpolated in gstat R package using two methods: the ordinary kriging and the multiple regression model that uses a set of terrain morphometric parameter as explanatory variables. Both methods produced an estimated precipitation map accompanied by a map of estimation error quantified with RMSE. The estimation by leave-one-out cross validation revealed that the multiple regression method produced much better accuracy, accounting for more than 90% of initial variance, compared with 63% for an ordinary kriging method. Combining and synthesizing both of these interpolation methods is possible with regression-kriging (being considered the best linear unbiased prediction model for spatial data). However, in this case it is not justified, as one of the methods performed considerably better than the other.

Keywords: climate modeling, precipitation, geostatistics, multiple regression, R.

Олександр Мкртчян

ОБРОБКА ТА ГЕОСТАТИСТИЧНА ІНТЕРПОЛЯЦІЯ ДАНИХ ЩОДО КІЛЬКОСТІ ОПАДІВ МЕТЕОСТАНЦІЙ ЗАХОДУ УКРАЇНИ

Метою дослідження була інтерполяція даних щодо середньорічних сум опадів, обрахованих шляхом обробки даних 50 метеорологічних станцій в західній частині України. Щоденні дані були отримані з відкритої бази даних GHCN. Шляхом їхньої обробки та сумування були виведені середньорічні суми опадів для кожної метеостанції. Опісля ці суми були проінтерпольовані в пакеті R gstat з використанням двох методів: звичайного кріґінґу та моделі множинної регресії, у якій в якості пояснювальних змінних використано морфометричні параметри рельєфу. Оцінка методом перехресної перевірки виявила, що другий метод приніс суттєво більш точний результат, який описує понад 90% варіабельності вихідних даних.

Ключові слова: моделювання клімату, опади, геостатистика, множинна регресія, R.

Александр Мкртчян

ОБРАБОТКА И ГЕОСТАТИСТИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ ДАННЫХ О КОЛИЧЕСТВЕ ОСАДКОВ МЕТЕОСТАНЦИЙ ЗАПАДА УКРАИНЫ

Целью исследования была интерполяция данных о среднегодовых суммах осадков, рассчитанных путем обработки данных 50 метеорологических станций в западной части Украины. Ежедневные данные были получены из открытой базы данных GHCN. Путем их обработки и суммирования были выведены среднегодовые суммы осадков для каждой метеостанции. Эти суммы затем были проинтерполированы в пакете R gstat с использованием двух методов: обычного кригинга и модели множественной регрессии, в которой в качестве объяснительных переменных использованы морфометрические параметры рельефа. Оценка методом перекрестной проверки выявила, что второй метод принес существенно более точный результат, который описывает более 90% вариабельности исходных данных.

Ключевые слова: моделирование климата, осадки, геостатистика, множественная регрессия, R.

Introduction. Precipitation amount is one of the most practically meaningful climatic elements that directly influences on the water resources and hydrologic processes, the conditions for agriculture and a number of other activities. There is a significant demand for accurate and reliable spatially distributed precipitation data. While precipitation maps are a customary component of thematic atlases, the scales of such maps are usually rather small (e. g. 1:8000000 for the maps in the National atlas of Ukraine [11]), and the methods of their creation are often vague and not formally defined, thus the accuracy and reliability of their information is unknown.

Modern tools for geospatial and statistical data analyses coupled with the availability of digital data sources make it possible to develop and apply formal and objective methods of mapping the climatic characteristics. As the primary source of climatic data the measurement data are collected on weather stations, the primary formal task connected with the creation of climatic maps is the one of the spatial interpolation of point data. Modeling of spatial fields through data interpolation is a common task of geostatistics. Simple geostatistical interpolation procedures like ordinary kriging based on the analysis of the spatial autocorrelation structure of the variable can be refined by considering the relationships of the variable of interest with other exhaustively-sampled explanatory variables available in the area of interest. In the case of climatic characteristics, these variables may correspond to the terrain morphometric parameters that influence the climate through their impact on energy balance and air masses movements.

The analysis of recent research. It is only quite recently that the geostatistical methods became commonly applied for the interpolation of climatic and specifically precipitation data. Phillips et al., performing interpolation of precipitation values in mountainous terrain in western Oregon, has found out that methods which take into account precipitation-elevation relationships, like detrended kriging and cokriging, provide better accuracy and precision compared with ordinary kriging [7]. In another study the mapping of the average precipitation from rainfall observations in a region of southern Italy has been performed using several methods, revealing that linear regression and ordinary cokriging has produced better results compared to the inverse distance interpolation while the best results (indicated by cross-validation) were produced by the multivariate geostatistical methods utilizing elevation data as an auxiliary explanatory variable [1]. In the study concerned with the mapping of monthly precipitation in Great Britain from sparse point data it was shown that kriging with an external drift (informed by elevation data) provided more accurate estimates than either ordinary kriging or deterministic moving window regression [6].

P. Goovaerts, comparing the performance of different interpolation techniques for the interpolation of monthly and annual precipitation data for 36 stations located in Southern Portugal, revealed that general linear regression of rainfall versus elevation gave much better predictions than methods which ignore elevation information (like inverse square distance method and ordinary kriging). However, the best results were obtained with methods that took into account elevation data while performing geostatistical interpolation [2]. The description of the general principles and theoretical foundations of geostatistical mapping, and the software tools for its realization together with some practical examples can be found in [3].

As of regional efforts to create spatially distributed climatic datasets, CARPATCLIM project should be mentioned that has been carried out by a consortium of institutions from nine countries with Hungarian Meteorological Service as the leading organization. Precipitation sums were among a set of examined variables that after quality-checking and homogenization were interpolated into 10-km resolution grids. The interpolation was implemented by MISH software that applies AURELHY method developed in 1980-th at the French Meteorological Service [8].

The purpose of this research is to create an accurate map of annual precipitation distribution with a well-documented and reproducible methodology using open data sources and software.

The main exposition. The area of interest encompasses the western part of Ukraine with the total area of ~156 thousand sq. km. Fifty meteorological stations where precipitation is regularly measured on standardized gauges are located inside the bounds of the study area. The data source used in the study consists of daily precipitation data from the Global Historical Climatology Network (GHCN) database [5], downloaded from the website of the European Climate Assessment & Dataset project http://www.ecad.eu/. The preprocessing of data involved several steps aimed at making data comparable and homogeneous and at summarizing data for further analyses. The downloaded data consisted of separate datafiles for every 50 meteorological station. R code has been written to remove headers and redundant columns, to merge separate data files into one data frame with columns that correspond to separate stations, to recode “nodata” values to “NA” R standard, and to split the DATE column into separate “year”, “month” and “day” parts.

The obtained data frame contained 33512 daily precipitation observations spanning from 1924 to 2011. No station, however, possessed the uninterrupted observation sequence for this time span, and periods of present and missing data values for different stations didn’t match. To solve the task, the decision was made to include into analysis only those observation dates for which the data was available for all 50 locations. Additionally, the time span was restricted to 1960–1990, when global climatic conditions were changing relatively slowly.

Thus, the total of 3432 daily observations have been selected. Monthly sums and averages were calculated taking care of different numbers of selected observations in different months. The locations of weather stations were mapped onto the shapefile that was imported into R spatial data frame using rgdal package, with the calculated annual precipitation sums merged into it by the common name column.

To prepare the data on terrain morphometric parameters, the 4 SRTM Version 4.1 DEM [4] tiles were downloaded (tiles 41_02, 41_03, 42_02, 42_03) and then merged into one raster. It has been then reprojected into UTM 35N coordinate system and resampled to 720 m resolution (this is justified by the low density of stations separated by tens of kilometers). A set of terrain morphometric parameters has been derived from DEM using focal (neighborhood) operators. These regard a terrain roughness factor, calculated as a variance of elevation values inside a circular moving window and aspect factor, calculated as differences in mean elevation values inside two opposite circular sectors. Is was hypothesized that increased terrain roughness could positively correlate with precipitation due to increased air flow turbulence that promotes vapor condensation, while the aspect influences precipitation values through well-known rain shadow and orographic precipitation effects. Each of these factors can be calculated on moving windows of different sizes, thus capturing the effects of different-scaled processes, and for the aspect factor different values of angles defining opposite circular sectors and corresponding cardinal points can be specified.

In our previous studies, different factors and scales combinations were tested for the strengths of their relationships with the precipitation sums for two separate years (1961, 1970) [9,10]. The terrain roughness factor was thus selected with 7.2 km moving window, while three different versions of aspect factor appeared to be independently significant, namely for the 36 km moving window (the aspect factors NW/SE and W/E), and for the 50.4 km moving window (the aspect factor NW/SE). The DEM data preprocessing and the derivation of morphometric parameters have been accomplished using appropriate SAGA GIS modules and gdal tools under QGIS.

The first part of the analysis aimed at the interpolation of data using ordinary kriging, that ignores the terrain morphometric parameters and interpolates data exploiting only the spatial distribution of precipitation values at data points. The R gstat package has been used for the task. To produce a sample (experimental) semivariogram, variogram () function was used. The vgm () function was used to specify a theoretical variogram, which requires specifying the values of variogram parameters and the selection of model type. Then fit.variogram () function was used to adjust the specified variogram parameters to better fit the data. Lastly, the theoretical variogram was used as a parameter to krige () function that carries out the interpolation, producing a spatial data frame that can be converted to raster object of either predicted values or prediction variance (rasterize function of raster package) and then to customary raster format like GeoTIFF. Krige.cv () function was used to perform cross-validation for kriging.

In our case, a well-pronounced empirical variogram has been produced, with nugget 30000 mm^2, partial sill 66000 mm^2, and range 468 km. The available theoretical variogram models have been tested by cross-validation, revealing that the best accuracy has been obtained by exponential model, following by spherical and Gaussian. The interpolated map visually lacks spatial detail and produces doubtful results for areas remote from nearest data points.

Another interpolation method applied was a multiple regression of precipitation values on terrain attributes mentioned above (fig. 1). To implement it, krige () function has been applied with the appropriate input formula and parameter model=NONE. The produced model shows pretty good fit (adjusted R-squared 0.9313, F-statistic 133.9 on 5 and 44 DF, p-value of model < 2.2e-16). Each of the terrain parameters used as predictors appeared to be statistically significant with $p < 0.01$. It appears that the terrain parameter having the strongest impact on precipitation is not the elevation ($t = 4.8$) but the terrain roughness ($t = 8.3$) which is in compliance with our former findings [9,10].

The precipitation map produced by regression model is characterized by much better spatial detail. Yet the best test of relative accuracy of two methods is given by cross-validation (Tab. 1). It appears that 36,2% of initial variance of values has been retained after the interpolation by ordinary kriging, while the regression model retains only 9.65 of initial data variability.

The validity of regression model has been assessed by examining the distribution of residual values, e.g. their normality being confirmed by Shapiro-Wilk test (shapiro.test () R function). Its obtained values (W = 0.986, p-value = 0.8125) suggest that the distribution of residual values is undistinguishable from normal. The model could be further refined by combining and synthesizing the two interpolation approaches into regression-kriging model that is considered the best linear unbiased prediction model for spatial data [3]. The gstat R package provides also for this most general interpolation technique. Yet in our case regression model residuals are not spatially correlated.

Table 1.
The initial and residual variance of precipitation values after application of two interpolation methods, estimated by leave-one-out cross validation

Initial	After ordinary kriging	After regression modeling
63096	22856 (36,2%)	6087(9,6%)

Conclusions and further prospects. The research demonstrates the opportunities that modern geostatistical methods provide for objective and reproducible analysis and spatial interpolation of point data of weather station records. Nowadays decent results can be obtained using open access data and software. The interpolation of precipitation values for 50 stations located in the western part of Ukraine showed that multiple regression of precipitation values on terrain morphometric attributes produced much better results compared with ordinary kriging. Further, the residuals of regression model showed no spatial autocorrelation, rendering unjustified in this case a more sophisticated regression-kriging method.

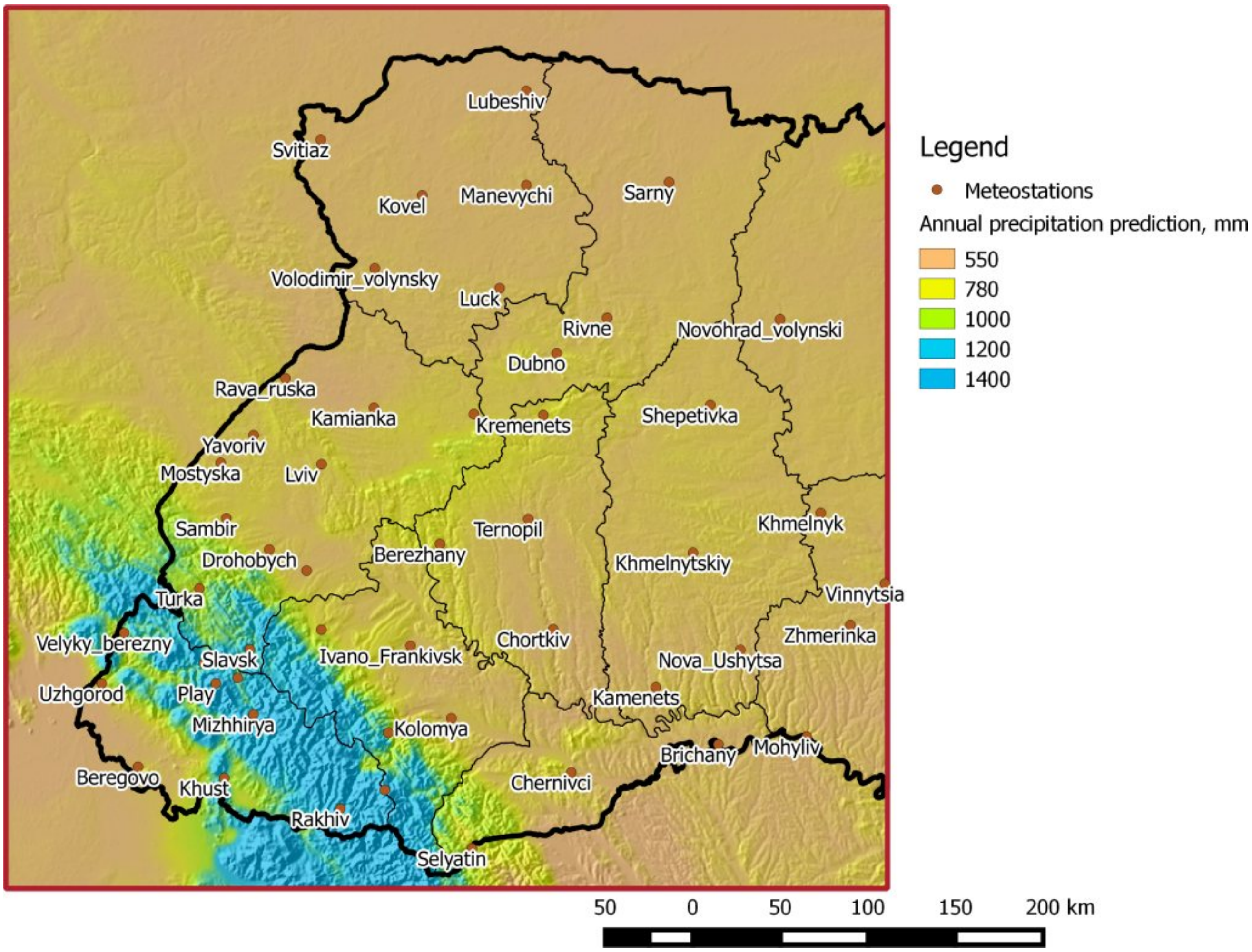

Fig. 1. Precipitation map produced by regression model

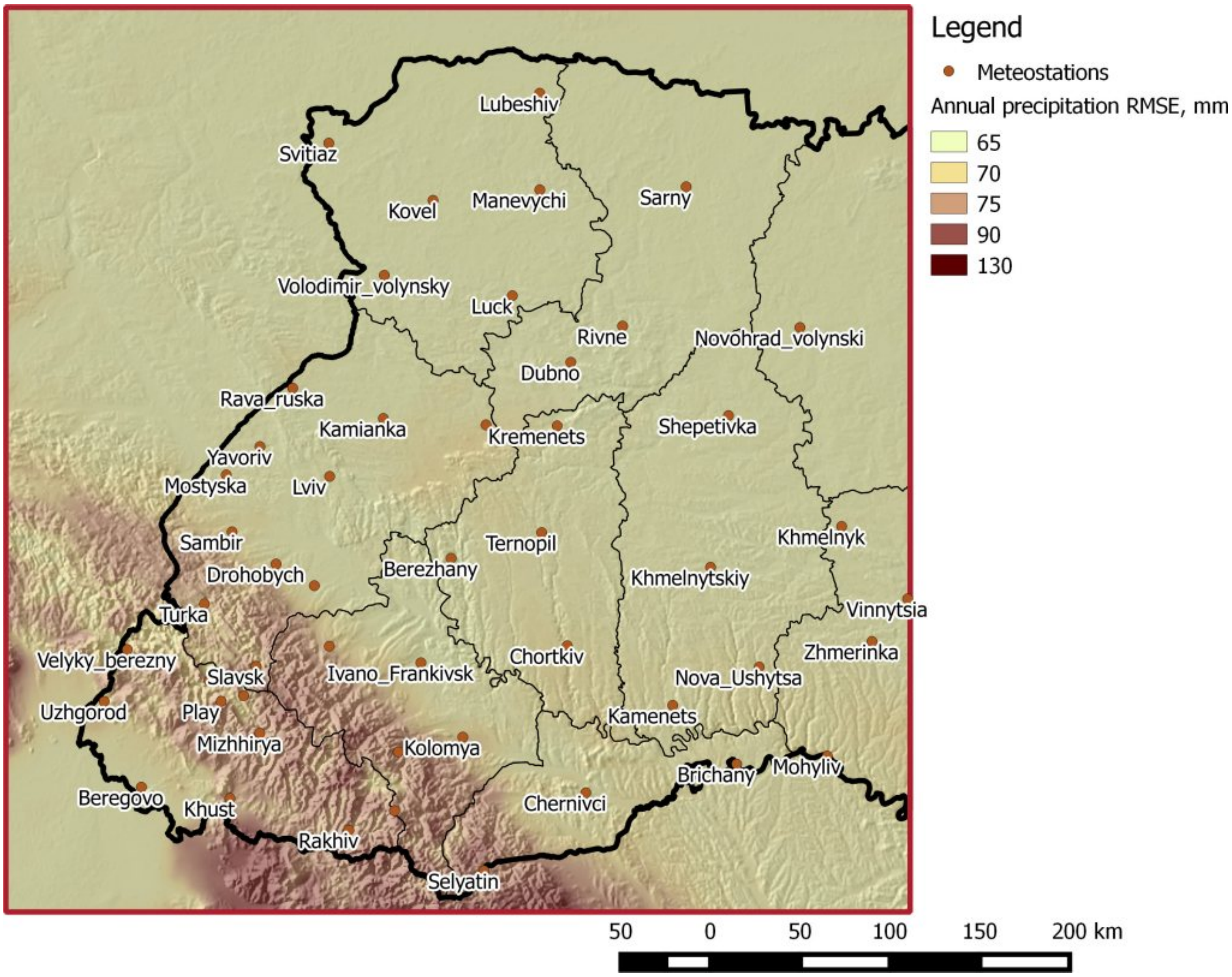

Fig. 2. Estimated RMSE of precipitation map produced by regression model

The spatial detail and accuracy of the interpolated result can be significantly increased by including the records of rain gauges, yet the problem arises of the their correct and precise georeferencing. Another commonly acknowledged problem is the undersampling of high elevation and complex terrain locations, making them underrepresented in factor space [7,8].

The resulting map of annual precipitation is reproducible due to the formal and transparent method of its creation. It is also accompanied by the map showing the spatial distribution of its estimated accuracy (Fig. 2).

The accuracy of interpolated results were compared with those of CARPATCLIM by calculating the REP parameter (see [8]) for a common subset of stations (those located in Ukrainian Carpathians) by leave-one-out cross-validation. The obtained value of 0.69 is slightly better than 0.66 of CARPACLIM precipitation prediction grid. While our modeling result cannot match the latter with its daily temporal detail and scores of predicted climatic variables, it has finer resolution (1 vs. 10 km) and no less accuracy, even while ignoring data records of rain gauges.

Reviewer: PhD in Geography, Associate Professor I.S. Kruglov

Література:

1. Diodato N. Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the Sannio Mountains (southern Italy) / N. Diodato, M. Ceccarelli // Earth Surface Processes and Landforms. – 2005. – Vol. 30, Iss. 3. – P. 259–268.

2. Goovaerts P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall / P. Goovaerts // Journal of Hydrology. – 2000. – Vol. 228. – P. 113–129.

3. Hengl T. A practical guide to geostatistical mapping / T. Hengl. – Amsterdam: University of Amsterdam, 2009. – 291 p.

4. Jarvis A. Hole-filled seamless SRTM data v4 [Електронний ресурс] / A. Jarvis, H. I. Reuter, A. Nelson, E. Guevara. – International Centre for Tropical Agriculture (CIAT), 2008. – Режим доступу : http://srtm.csi.cgiar.org.

5. Klein Tank A.M.G. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment / Klein Tank A.M.G. and Coauthors // Int. J. of Climatol. – 2002. – Vol. 22. – P. 1441-1453.

6. Lloyd C. D. Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain / C. D. Lloyd // Journal of Hydrology. – 2005. – Vol. 308 (1–4). – P. 128–150.

7. Phillips D. L. A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain / D. L. Phillips, J. Dolph, D. Marks // Agric. For. Meteorol. – 1992. – Vol. 58. – P. 119–141.

8. Spinoni J. Climate of the Carpathian Region in 1961-2010: Climatologies and Trends of Ten Variables / J. Spinoni and the CARPATCLIM project team (39 authors) // International Journal of Climatology. – 2015. – Vol. 35, Iss. 7. – P. 1322–1341.

9. Мкртчян О. Інтерполяція даних метеоспостережень кількостей опадів та інших кліматичних змінних методом регресійного криґінґу / О. Мкртчян, П. Шубер // Вісник Львів. ун-ту. Сер. геогр. – 2013. – Вип. 42. – С. 258-264.

10. Мкртчян О. Методика геопросторового моделювання та картування кліматичних характеристик за даними спостережень метеостанцій / О. Мкртчян, П. Шубер // Вісник Львів. ун-ту. Сер. геогр. – 2011. – Вип. 39. – С. 245-253.

11. Національний атлас України. – К.: ДНВП “Картографія”, 2007.

References:

1. Diodato, N., Ceccarelli, M. (2005). Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the Sannio Mountains (southern Italy). Earth Surface Processes and Landforms, 30 (3), 259–268.

2. Goovaerts, P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, 228, 113–129.

3. Hengl, T. (2009). A practical guide to geostatistical mapping. Amsterdam: University of Amsterdam, 291.

4. Jarvis, A., Reuter, H. I., Nelson, A., Guevara, E. (2008). Hole-filled seamless SRTM data v4. International Centre for Tropical Agriculture (CIAT). Available at: http://srtm.csi.cgiar.org.

5. Klein Tank, A. M. G., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol., 22, 1441-1453.

6. Lloyd, C. D. (2005). Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. Journal of Hydrology, 308 (1–4), 128–150.

7. Phillips, D. L., Dolph, J., Marks, D. (1992). A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain Agric. For. Meteorol, 58, 119–141.

8. Spinoni, J., the CARPATCLIM project team (39 authors). (2015). Climate of the Carpathian Region in 1961-2010: Climatologies and Trends of Ten Variables International Journal of Climatology, 35 (7), 1322–1341.

9. Mkrtchian, A., Shuber, P. (2013). Interpolyaciya dany`x meteosposterezhen` kil`kostej opadiv ta inshy`x klimaty`chny`x zminny`x metodom regresijnogo kry`g`ing`u [Interpolation of weather observation data on precipitation and other climatic variables by the regression kriging method]. Visnyk of Lviv University. Ser. geogr., 42, 258-264.

10. Mkrtchian, A., Shuber, P. (2011). Metody`ka geoprostorovogo modelyuvannya ta kartuvannya klimaty`chny`x xaraktery`sty`k za dany`my` sposterezhen` meteostancij [Methods of geospational modeling and mapping of climatic characteristics from the meteorologic stations records]. Visnyk of Lviv University. Ser. geogr., 39, 245-253.

11. Nacional`ny`j atlas Ukrayiny` [National atlas of Ukraine]. (2007). Ky`yiv : DNVP “Kartografiya”.