Rev. Bras. Agrometeorologia, v. 14, n.2, p.225-233, 2006
Recebido para publicação em 08/02/06. Aprovado em 10/08/06.

ISSN 0104-1347

Filling in missing rainfall data in the Andes region of Venezuela, based on a cluster analysis approach

Preenchimento de falhas em séries de precipitação pluvial na região dos Andes, Venezuela, baseado na análise de agrupamento

Beatriz Ibet Lozada Garcia[1], Paulo Cesar Sentelhas[2], Luciano Tapia[3], Gerd Sparovek[4]

Abstract: Several agrometeorological studies require daily rainfall, mainly those which have as objective to model or simulate water budget, crop development and yield, and occurrence of crop pests and diseases. However, the presence of missing data is a problem that normally occurs, which limits these studies. The origin of this problem varies, but they are more related to improperly devices functioning or lack of technical officers to make observations. Such problems are common in the different institutions which control weather-stations networks in Venezuela. Simple and feasible alternatives to improve the quality of rainfall database are required. We established the hypothesis that daily rainfall data from a weather station can be used to fill in missing data from another surrounding weather station. Data used to test our hypothesis were obtained from 106 weather stations in the Andes region, Venezuela, considering a period of 31 years (1967-1997). The original rainfall database presented 17.3% of missing data (207,534 days). Using a cluster analysis (Ward's method, with Euclidean distance), the proposed method, named as *Closest Station*, reduced the percentage of missing data to 2.5% (29,495 days). The performance of our proposed method was evaluated by mean absolute error (MAE), which ranged from 1.7 to 4.0 mm day^{-1}, and by Willmott agreement index (d), which was 0.57 for daily basis and 0.83 for monthly basis. The contingency analysis showed that our proposed method overestimated rainfall events for daily data, which resulted in a smaller fraction of correct estimates (FC = 0.48) and a larger false alarm ratio (FAR = 0.53), limiting their use. For the other time scales, from 7 to 30 days, FC was greater than 0.88 and FAR smaller than 0.07, which allow the use of this technique for several purposes in agrometeorological studies.

Key-words: *Closest Station* method, precipitation, climatology, database

Resumo: Vários estudos agrometeorológicos exigem o uso de dados diários de chuva, especialmente os que têm como objetivo a modelagem e simulação do balanço hídrico, do crescimento, desenvolvimento e rendimento das culturas, da ocorrência e proliferação de pragas e doenças. Entretanto, a presença de dados faltantes nas séries de dados é um problema que normalmente ocorre, limitando assim tais estudos. As origens de tais problemas são diversas, mas estão principalmente relacionadas ao mau funcionamento dos equipamentos e à falta de observadores. Tais problemas são muito comuns nas redes pluviométricas das mais diversas instituições venezuelanas. Sendo assim, há a necessidade de se desenvolver técnicas simples e factíveis, que possibilitem o preenchimento das falhas existentes nas séries históricas, melhorando a qualidade dos bancos de dados de chuva. Baseado nisso, estabeleceu-se a hipótese de que os dados diários de chuva de uma dada estação podem ser usados para preencher as falhas de uma estação vizinha próxima. Os dados utilizados para testar a hipótese formulada foram obtidos de 106 estações meteorológicas da região dos Andes, Venezuela,

[1] Instituto Nacional de Investigaciones Agrícolas (INIA), Bramón, Táchira, Venezuela. E-mail: blozada@inia.gov.ve
[2] Departamento de Ciências Exatas – ESALQ, Universidade de São Paulo, Piracicaba, Brasil. E-mail: pcsentel@esalq.usp.br
[3] Centro de Informática na Agricultura (CIAGRI) - Universidade de São Paulo, Piracicaba, SP, Brasil. E-mail: lrtapia@esalq.usp.br
[4] Departamento de Ciência do Solo – ESALQ, Universidade de São Paulo, Piracicaba, SP, Brasil. E-mail: gerd@esalq.usp.br

considerando-se um período de 31 anos (1967-1997). O banco de dados original apresentava 17,3% de dados faltantes (207.534 dias). Utilizando-se a análise de agrupamento (método de Ward com distância Euclidiana), o método proposto, denominado de *Estação mais Próxima*, reduziu a porcentagem de dados faltantes para 2,5% (29.495 dias). O desempenho do modelo proposto foi avaliado pelo erro absoluto médio (MAE), que variou de 1,7 a 4,0 mm dia^{-1}, e pelo índice de concordância de Willmott (d), que foi de 0,57 para a escala diária e de 0,83 para a escala mensal. A análise de contingência mostrou que o método da *Estação mais Próxima* superestimou os eventos de chuva na escala diária, o que resultou em uma menor fração de estimativas corretas (FC = 0,48) e em uma maior taxa de estimativas erradas (FAR = 0,53), limitando, assim, seu uso. Para as outras escalas temporais, de 7 a 30 dias, FC foi maior que 0,88 e FAR menor que 0,07, o que permite o uso dessa técnica para o preenchimento de falhas em séries de dados a serem utilizadas em estudos agrometeorológicos.

Palavras-chave: Método da *Estação mais Próxima*, chuva, climatologia, séries históricas.

Introduction

A complete and accurate source of rainfall data is required for an efficient modeling of a wide variety of environmental processes (JEFFREY et al., 2001). According to WILKS (1999), an important limiting factor in application of agricultural, hydrological and ecosystem models can be the lack of weather data with sufficient temporal and spatial coverage. In agrometeorology, these processes and models are mainly related to water budget, crop development and yield, and crop pests and diseases occurrence. Often, agrometeorologists face the problem of missing data due to a variety of reasons, which restrict the research efforts of many workers. In general, the reasons for missing data occurrence are related to: records for discrete periods, not covering the entire time period of interest; short intermittent periods where data have not been recorded; and systematic or random errors (PECK, 1997). Such problems are common in the different institutions which control weather-stations networks in Venezuela, making the use of rainfall data limited for studies in climatology, agrometeorology, and hydrology.

As solution for this problem, several techniques have been proposed. These techniques for estimating missing weather data can be grouped in: empirical methods, statistical methods, and function fitting (THIEBAUX & PEDDER, 1987). The empirical approaches include methods like simple arithmetic averaging, inverse distance interpolation, and the closest station. The statistical techniques use multiple regression analysis, multiple discriminant analysis, principal component analysis, cluster analysis, kriging technique and optimal interpolation. And in function fitting, data are fitted as a function like thin-plate spline, which now has been used to interpolate the climatological data (XIA et al., 1999; PRICE et al., 2000; HASENAUER et al., 2003; TEEGAVARAPU & CHANDRAMOULI, 2005). The use of remotely sensed data has also been an option to filling in missing data (JEFFREY et al., 2001), despite the difficulties related to its implementation.

Cluster analysis is one of the statistical techniques often used in meteorology and climatology to identify homogeneous climate groups and for climate classification (GERSTENGARBE et al., 1999; DEGAETANO, 2001). The aim of the cluster analysis is the separation of several elements into homogeneous groups where weather data in such period of time can be considered similar. It can be performed by hierarchical and non-hierarchical techniques. However, hierarchical clustering methods are ideal for the exploratory stage of research and the most common used for climate research is Ward's minimum variance (UNAL et al., 2003). To search for the most similar pairs to merge, cluster analysis requires specific measurements of dissimilarity to characterize the relationships among the stations. The most common distance metric used in meteorology/climatology studies is Euclidean distance. All the hierarchical methods follow the basic four-step routine to identify those sublets that are both homogeneous and heterogeneous: 1) calculation of the specified distance measure between all entries (weather data); 2) formation of a new cluster merging from the two closest entries, based on a defined criterion; 3) recalculation of the distance between all entries; and 4) repetition of steps 2 and 3 until all entries are merged into one cluster

According to DEGAETANO (2001), one problem of the cluster analysis is to determine an

Rev. Bras. Agrometeorologia, v. 14, n.2, p.225-233, 2006

adequate true number of clusters, since the method continues to merge station groups until ultimately one cluster encompassing all sites. WILKS (1999) and GERSTENGARBE et al. (1999) presented some statistical rules to guide the termination of cluster mergers, but even these tend to be somewhat subjective. Thus, adequate number of clusters should be determined by the level of detail desired in a particular study.

Considering the closest station empirical method (XIA et al., 1999) and the characteristics of cluster analysis (UNAL et al., 2003), we established the hypothesis that daily rainfall data from a weather station can be used to fill in missing data from another surrounding weather station, since they were considered similar by the cluster analysis, using Ward's method, with Euclidian distance. For that, we set the following goals:

a) Determine the two closest stations for each one of the 106 weather stations used in this study, considering cluster analysis;

b) Fill in missing rainfall data with those from the closest stations;

c) Evaluate the performance of our proposed method considering 1,000 rainy periods for daily, weekly, bi-weekly, and monthly time scales.

Material and Methods

Location and rainfall data

Our study is based on daily rainfall data obtained from 106 stations, belonged to Ministry of Environment and Natural Resources (MARN) of Venezuela, distributed in three different states of the Andes region: Táchira, Mérida and Trujillo, with altitude ranging from 0 to 5,000 meters (Figure 1), for a period of 31 years, from 1967 to 1997. Missing rainfall data existed at all stations due mainly to interruptions in observations. A total of 1,199,390 days of missing data, about 17% of the total, were detected after a consistence analysis.

Cluster analysis

Cluster analysis was applied to identify the two most similar stations for each one of the 106 rainfall stations, considering each state separately and the following variables: monthly rainfall, latitude, longitude, and altitude. For this analysis, the hierarchical Ward's method was used, with the dissimilarity measure given by the Euclidean distance (WARD, 1963):

$$d_{ij} = \frac{N}{M}\sqrt{\sum_{k=1}^{M}\left(x_{ik} - x_{jk}\right)^2} \qquad (1)$$

where d_{ij} is the Euclidean distance between x_i and x_j over M available data points. N is the number of data points for the whole period. Each variable was standardized prior to distance calculation to eliminate the scale effect, since observation with different scales may unequally contribute to the calculated distance, and the distances between stations were standardized by using the common period and full period of the database. The threshold considered to identify a station similar to another varied from 100 to 500 depending on the month.

The first closest station was considered up to the limit of 100 and the second up to 500. It means that such a given station just has a similar one if the link between them is below the limit of 100 for the first option and 500 for the second option, as presented in the example of Figure 2.

The results from cluster analysis were used to build a table with the list of the 106 rainfall stations and their two closest stations.

Filling in missing rainfall data

After the closest stations were identified by cluster analysis, the missing rainfall data were estimated from the data of the first closest station. If the first closest station also presented missing data, the estimates were done with rainfall data from the second closest station. When both closest stations presented missing data, the main station remained without data for such period of time. Figure 3 presents the flux diagram of the procedures to fill in missing rainfall data. All these steps to filling in missing rainfall data were done with a computer program, developed in Visual Basic 6.0. This program uses the table with the closest stations for each one of the 106 rainfall stations, determined by cluster analysis. Knowing the 1st and 2nd closest stations, the program starts its routine by checking

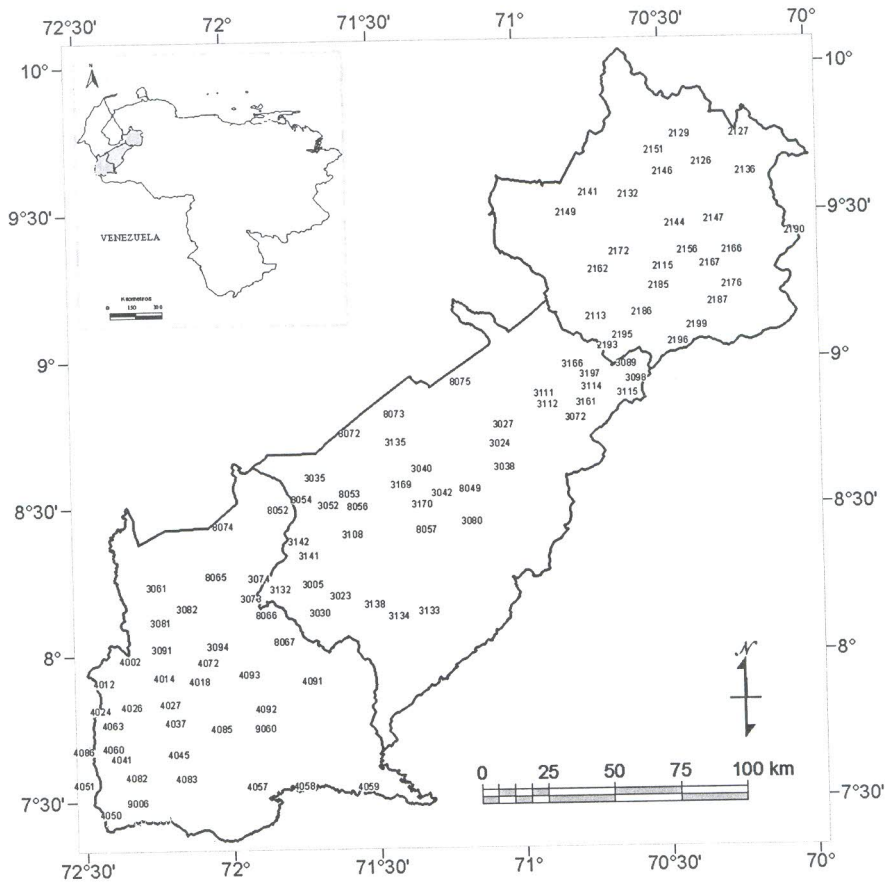

Figure 1. Location and reference number of the rainfall stations, distributed in the states of Táchira, Mérida and Trujillo, Venezuela.

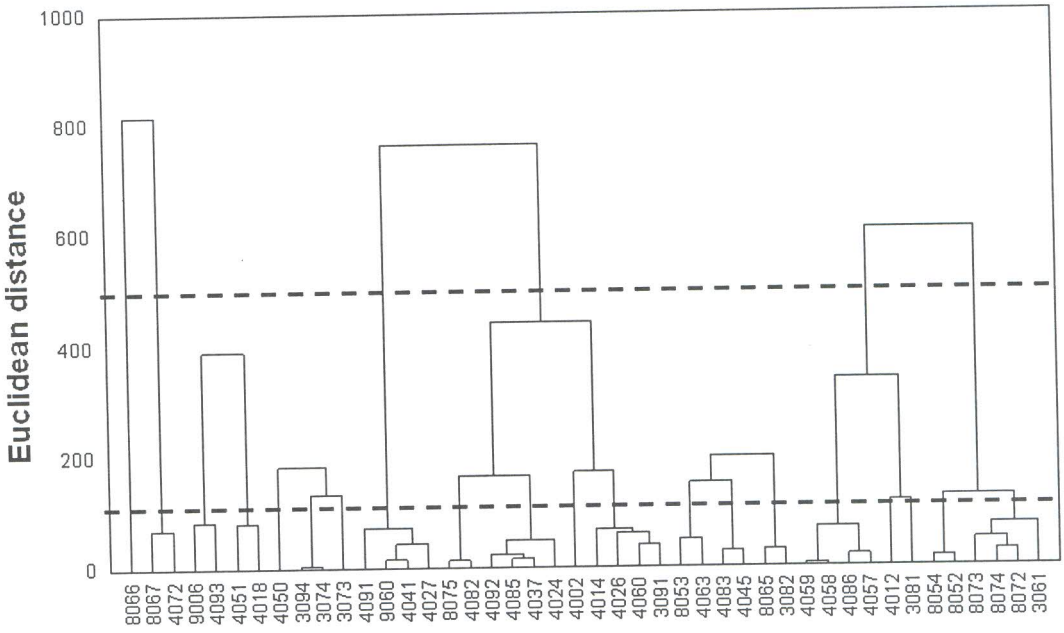

Figure 2. An example of a diagram showing the linkages between rainfall stations, considering the results from cluster analysis for a given month. Dashed lines represent the distance threshold used to consider similar stations (first option, up to 100; second option, up to 500).

Rev. Bras. Agrometeorologia, v. 14, n.2, p.225-233, 2006

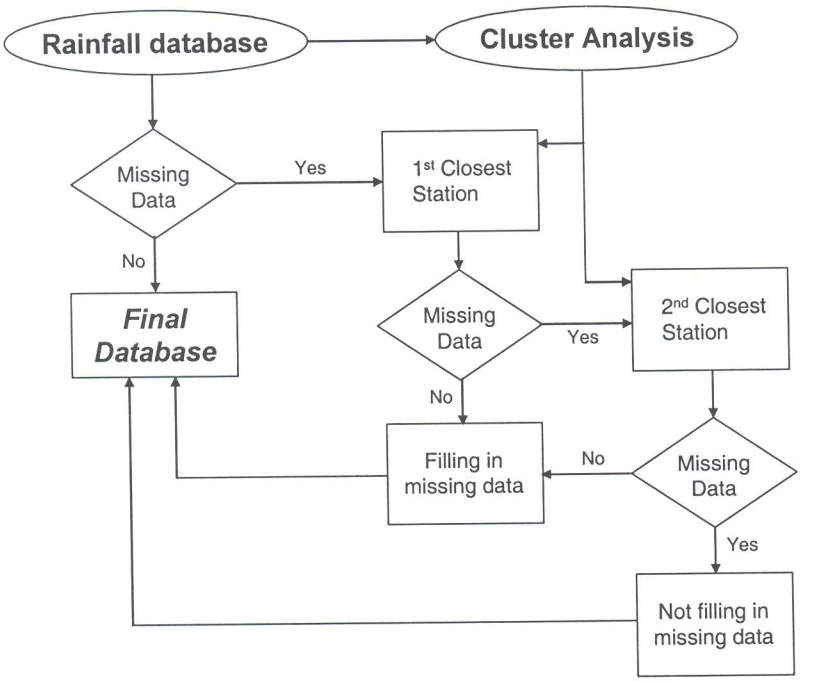

Figure 3. Flux diagram showing the procedures for filling in missing rainfall data, based on the cluster analysis approach, named *Closest Station* method.

each day of the database in relation to missing data. When a day with missing data is identified, the program looks for the 1st closest station and checks if this station has missing data for such day – if *no*, the program fill in the missing data and send this data to the final database, if *yes*, the program looks for the 2nd closest station. When the 2nd closest station is identified, the program checks if this station has missing data for such day - if *no*, the program fill in the missing data and send this data to the final database, if *yes*, this day remains without rainfall data.

Performance of the proposed method to fill in missing rainfall data

To check our method, rainfall data from 1,000 periods of time were taken randomly from the database of the 106 stations, which were organized in four different time scales: daily, weekly, bi-weekly, and monthly. These data were taken out from the data set and estimated using our *Closest Station* method. After that, observed and estimated rainfall data were compared and the following indexes and errors were determined (WILLMOTT, 1981; ZACHARIAS et al., 1996):

a) Root Square Mean Error (RSME)

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2\right]^{0,5} \quad (2)$$

b) Willmott Agreement Index (d)

$$d = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \quad (3)$$

c) Mean Error (ME)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \quad (4)$$

d) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (5)$$

where: O_i and P_i are measured and estimated rainfall data respectively, n is the number of data, and $\overline{O}$ is the average of measured data. Correlation coefficient (r) was also used to evaluate the performance of the method.

Our method was also evaluated considering three statistical scores: Fraction of Correct Estimates (FC), which is the percentage of days with correct estimations; False Alarm Ratio (FAR), which is the percentage of days with wrong rainfall estimations; and Bias, which is the ratio between the total of estimated and measured rainy periods. If Bias is greater than 1, the number of rainfall events was overestimated. If Bias is smaller than 1, it means that the method underestimated the rainfall periods of time. So, a method with a good performance should have a Bias close to 1. To calculate these scores, we used a contingency table presented by WILKS (1995) (Table 1).

If the method correctly estimates rainfall for such day, it is accounted in box *A*. If the method does not estimate rainfall for such day, while it does occur, it is accounted in box *B* (misses). If the method estimates rainfall for such day, while it does not occur, it is accounted in box *C* (false prediction). And if the method correctly does not estimate rainfall for such day, it is accounted in box *D*. Using the contingency table, FC, FAR, and Bias were calculated with the following equations (WILKS, 1995):

$$FC = \frac{(A+D)}{n} \quad (6)$$

$$FAR = \frac{C}{(A+C)} \quad (7)$$

$$Bias = \frac{(A+C)}{(A+B)} \quad (8)$$

where n is the number of periods of time in the respective time scale (= 1,000).

Table 1. Contingency table for calculation of statistical scores, considering measured and estimated rainfall data.

	Observed	
Estimated	Yes	No
Yes	A	B
No	C	D

Source: WILKS (1995).

Results and Discussion

The Cluster Analysis allowed determining similar stations in relation to rainfall data and the two closest stations for each station were identified. Among the 106 stations considered in this study, just two, one in the state of Táchira and the other in the state of Mérida, did not have a second closest station. The two stations are close to the west border of the Andes region and at very low altitude (220 m). In these two stations, missing data were filled in just with the first closest station.

Applying our proposed method, named the *Closest Station*, for daily data, the percentage of missing data was reduced from 17.3% (207,534 days) to 2.5% (29,495 days), considering that the original rainfall database has 1,199,390 days (31 years). This is an impressive reduction in the missing data, mainly if we considered that our method just uses the two closest stations determined by cluster analysis. XIA et al. (1999), using a closest station method, based only on the distance between stations, filled in all the missing data for a 4-year period in Germany. However, these authors considered the five closest stations, using 44 weather stations in Bavarian region. Similar results were found by TEEGAVARAPU & CHANDRAMOULI (2005) in Kentucky, USA, with 4 closest rainfall stations. In this case, the closest stations were identified by assessing the similarity in the geometric patterns of the observed rainfall time series.

The performance of the *Closest Station* method is given in Table 2. When the method was used to fill in missing data in 1,000 periods in different time scales, a slight tendency of overestimation was observed for daily data (ME = +0.157 $mmday^{-1}$) and of underestimation for the other time scales (ME < 0). Values of MAE decreased when the time scale increased, which is also observed in RSME. Similar results were presented by XIA et al. (1999), what according to JEFFREY et al. (2001) would be expected since daily rainfall has a spatial variability much larger than monthly rainfall.

Rev. Bras. Agrometeorologia, v. 14, n.2, p.225-233, 2006

Table 2. Mean error (ME), mean absolute error (MAE), root square mean error (RSME), coefficient of correlation (r), and Willmott agreement index (d) for the comparison between observed and estimated rainfall at Andes region of Venezuela.

Time scale	N	ME	MAE	RSME	r	d
			(mm day^{-1})			
Daily	1000	+0.157	4.042	9.075	0.345	0.565
Weekly	1000	-0.200	2.548	4.119	0.531	0.723
Bi-weekly	1000	-0.868	2.340	3.548	0.556	0.741
Monthly	1000	-2.589	1.684	2.577	0.695	0.830

n = number of periods considered in the analysis.

MAE and RSME values obtained in this study are slightly greater than those presented by XIA et al. (1999) for Bavaria, Germany, and TEEGAVARAPU & CHANDRAMOULI (2005), for Kentucky, USA, when using similar methods. These authors found MAE ranging from less than 1 mm day^{-1} to 3 mm day^{-1}, however, considering a larger number of closest stations and a smaller geographical area.

Considering now the correlation between observed and estimated rainfall data, it is clear that our proposed method had a moderate performance for daily scale, with $r = 0.34$ and $d = 0.56$ (Table 2). On the other hand, when longer periods were analyzed a better performance was observed, with $r > 0.53$ and $d > 0.72$, allowing the use of these estimates in models where rainfall data are required. Another way to evaluate the performance of *Closest Station* method for filling in missing rainfall data is by statistical scores, calculated from the contingency table (Table 3). Table 4 presents the statistical scores - fraction of correct estimates (FC), false alarm ratio (FAR), which only give the proportion of events correctly and wrongly estimated, and Bias, which describes the direction of the error.

From Table 3 we can see that *Closest Station* method was also moderate to estimate rainfall events in a daily scale, which could be the main source of rainfall amount errors, as presented in Table 2. For daily scale, the method just correctly predicted rainfall in 33% of days, while for the other time scales the proportion of correctly predicted rainfall events was greater than 0.8.

Table 3. Contingency table for the *Closest Station* method, considering the different time scales. Values are given in fraction of 1,000 days.

Daily		
	Estimated	
Observed	Rain	No rain
Rain	0.33	0.15
No rain	0.37	0.15

Weekly		
	Estimated	
Observed	Rain	No rain
Rain	0.80	0.06
No rain	0.06	0.08

Bi-Weekly		
	Estimated	
Observed	Rain	No rain
Rain	0.92	0.03
No rain	0.02	0.03

Monthly		
	Estimated	
Observed	Rain	No rain
Rain	0.97	0.02
No rain	0.00	0.01

Table 4. Statistical scores for the *Closest Station* method, considering the different time scales. Values are given in proportion in relation to 1,000 events

Time scale	FC	FAR	Bias
Daily	0.48	0.53	1.46
Weekly	0.88	0.07	1.00
Bi-weekly	0.95	0.02	0.99
Monthly	0.98	0.00	0.98

The statistical scores presented in Table 4 can express better the performance of the method to estimate rainfall periods at different time scales. In a daily scale, our method only predicted correctly events, with or without rain, in 48% of the days (FC = 0.48), against 53% with false alarm (FAR = 0.53), or better with predicted rain when there was no rain or with predicted no rain when there was rainfall. Again, the Bias = 1.46 is showing that in this time scale there was an overestimation of the rainfall periods, which was responsible by the overestimation of rainfall amounts (Table 2). For the other time scales, the performance of the model improved gradually, with FC ranging from 0.88 to 0.98. On the other hand, FAR decreased to close to zero, and Bias stayed around 1, agreeing with Table 2 and showing the good performance of *Closest Station* method to estimate rainfall events and amounts for periods of 7, 15 and 30 days.

Conclusions

The use of *Closest Station* method, proposed in this study, reduced the missing rainfall records in the database by about 85%. For daily data, our method presented a moderate performance, failing for the most rainfall events and amounts. For the other time scales studied (7, 15 and 30 days), it showed a very good performance, similar to those presented by other authors. The *Closest Station* method, with two closest stations determined by cluster analysis, could be used to estimate the missing rainfall data for different time-scales at the Andes region of Venezuela.

References

DEGAETANO, A. Spatial grouping of United States climate stations using a hybrid clustering approach. **International Journal of Climatology**, Chichester, v. 21, p.791-807, 2001.

GERSTENGARBE, F.-W.; WERNER, P.C.; FRAEDRICH, K. Applying non-hierarchical cluster analysis algorithms to climate classification: some problems and their solution. **Theoretical and Applied Climatology**, Wien, v.64, p.143-150, 1999.

HASENAUER, H. et al. Validating daily climate interpolations over complex terrain in Austria. **Agricultural and Forest Meteorology**, Amsterdam, v. 119, p.87-107, 2003.

JEFFREY, S.J. et al. Using spatial interpolation to construct a comprehensive archive of Australian climate data. **Environmental Modeling & Software**, Amsterdam, v. 16, p.309-330, 2001.

PECK, E.L. Quality of hydrometeorological data in cold regions. **Journal of the American Water Resources Association**, Middleburg, v.33, p.125-134, 1997.

PRICE, D.T. et al. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. **Agricultural and Forest Meteorology**, Amsterdam, v.101, p.81-94, 2000.

TEEGAVARAPU, R.S.V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. **Journal of Hydrology**, Amsterdam, v.312, n.1-4, p.191-206, 2005.

THIEBAUX, H.J.; PEDDER, M.A. **Spatial objective analysis with applications in atmospheric sciences**. London: Academic Press. 1987. 455p.

Rev. Bras. Agrometeorologia, v. 14, n.2, p.225-233, 2006

UNAL, Y.; KLINDAP, T.; KARACA, M. Redefinig the climate zones of Turkey using cluster analysis. **International Journal of Climatology**, Chichester, v.23, n.9, p.1045-1055, 2003.

WARD, J.H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v.58, n.1, p.236-244, 1963.

WILKS, D. **Statistical methods in the atmospheric sciences**. New York: Academic Press, 1995. 467p.

WILKS, D. Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. **Agricultural and Forest Meteorology**, Amsterdam, v.96, p.85-101, 1999.

WILLMOTT, C.J. On the validation of models. **Physical Geography**, Norwich, v.2, n.2, p.184-194, 1981.

XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: estimation of missing values for Bavaria, Germany. **Agricultural and Forest Meteorology**, Amsterdam, v.96, p.131-144, 1999.

ZACHARIAS, S.; HEATWOLE, C.D.; COAKLEY, C.W. Robust quantitative techniques for validating pesticide transport models. **Transactions of the ASAE**, St. Joseph, v.39, n.1, p.47-54, 1996.