A PEER-TO-PEER COMPUTING SYSTEM FOR GAP FILLING IN CLIMATE RECORDS

Luiz Rafael Schmitke, Márcio Augusto de Souza, Jorim Souza das Virgens Filho, Luciano José Senger[1]

[1] Universidade Estadual de Ponta Grossa, Ponta Grossa, PR, Carlos Cavalcanti Av., 4748, 84030-900, phone +55 42 3220 3300, Brazil. E-mails: luizrafael88@gmail.com, msouza@uepg.br, jvirgens@uepg.br, ljsenger@uepg.br.

Abstract. *This work presents a computing model for filling gaps in climate records obtained from automatic weather stations. A peer-to-peer distributed computing system is used for sharing climate records among the stations and a geographic criterion is adopted for selecting climate records from nearby stations to fill gaps in records. The model was evaluated using real climate records and a simulated data set, with gaps in the series, as inputs. The results obtained were compared to real and to simulation-generated data for estimating the quality of the data generated by the proposed model. The computing model achieved accurate results for the months considered in the case study and it was able to generate climate data that can statistically be considered as belonging to the same population of the real climate data studied.*

Keywords. *distributed systems, climate data, peer-to-peer computing.*

1. INTRODUCTION

Agriculture is an economic activity that suffers great influence from the climate. According Hoogenboom [1], a great part of the variation in agricultural productivity is due to natural factors that cannot be controlled.

Although controlling the climate is not possible, farmers and researchers can use tools to predict and simulate the behavior of climate in future dates, minimizing its impact. These tools make its prediction based on climatic data collected at conventional (CWE) or automatic (AWE) weather stations.

AWEs can be affected by problems of signal interference, disconnection and cable oxidation, which can generate abnormal data or gaps in climate records. According to Mateo and Leung [2] and Hoogenboom [1], the accuracy of real climate records have great importance for research in agriculture, but it is also important to develop methods to correct gaps and/or abnormalities in these records. For this purpose, statistical methods are adopted for generating and correcting climate data series. For instance, Stochastic models are capable of generating synthetic climate data from the real climate data [7,8,9].

This work presents a system capable of correcting gaps in climate records, which uses climate values obtained in neighboring regions for filling the gaps in a particular region. This system is organized as a peer-to-peer network, which does not has a central server and is formed by computers, called peers, that are independent and can act as servers or clients.

This paper also shows a case study of the use of the system, and the results obtained are compared to real and to simulation-generated data for estimating the quality of the data generated by the proposed system.

2. METHODOLOGY

The climate data series used in the case study were obtained from SIMEPAR and ABC Foundation (FABC), and it contains data collected in daily basis from AWEs in locations shown in Table 1.

City	Latitude	Longitud e	Period	Provider
Ponta Grossa	-25,05	-50,09	1997-2011	SIMEPAR
Telêmaco Borba	-24,20	-50,37	1997-2011	SIMEPAR
Fernandes Pinheiro	-25,25	-50,32	1997-2011	SIMEPAR
Castro	-24,79	-50	2008-2012	FABC
Tibagi	-24,53	-50,37	2009-2012	FABC
Carambeí	-24,87	-50,22	2009-2012	FABC
Piraí do Sul	-24,40	-50,10	2009-2012	FABC

Table 1 – Location of weather stations

It was used the climate information obtained in *Carambeí* to run the tests of the correction model, because it has a central position in relation to the other cities, as show in Figure 1. In the climate data collected in *Carambeí*, gaps were artificially generated in the periods presented in Table 2.

Month	Period	Days
February	01 a 10	10
February	11 a 20	10
February	19 a 28	10
February	01 a 20	20
February	09 a 28	20
February	01 a 28	28
May	01 a 10	10
May	11 a 20	10
May	22 a 31	10
May	01 a 20	20
May	12 a 31	20
May	01 a 31	31

August	01 a 10	10
August	11 a 20	10
August	22 a 31	10
August	01 a 20	20
August	12 a 31	20
August	01 a 31	31
November	01 a 10	10
November	11 a 20	10
November	21 a 30	10
November	01 a 20	20
November	11 a 30	20
November	01 a 30	30

Table 2 – Periods in which gaps are generated

The experiments were applied to values of temperature, which is a climate variable whose values have a normal distribution. According to Steinnhaeuser et al. [6], Sentelhas and Monteiro [5] and Hoogenboom [1], temperature is the most significant climate variable when it is desired to determine climatic regions or indexes that act as predictors.

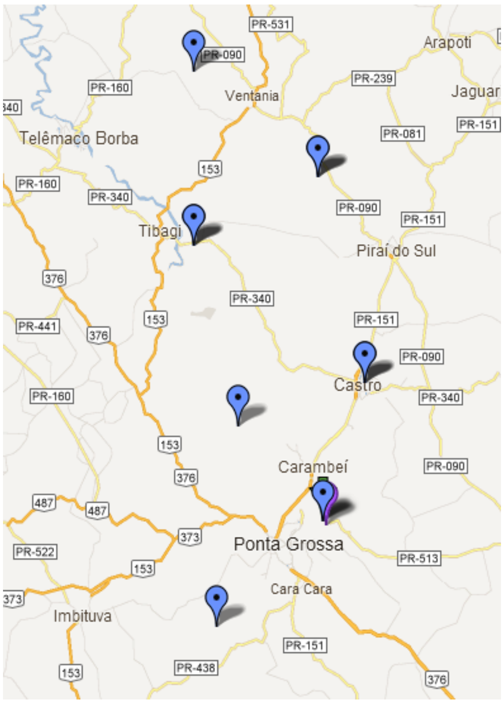

Figure 1 – Map of weather stations

The proposed correction system was developed using Java programming language, and it was executed on a P2P network built using P2PComp framework, proposed by Senger et al. [4]. The P2PComp framework defines an infrastructure for transparently starting and monitoring parallel applications written with Java. Additionally, it also provides a programming library to be included in the code of the programs executed by the framework.

In the proposed system, each P2P peer is responsible for climate data of a weather station. The correcting gaps process is carried out by two peers, which are located in all computers of the P2P network. The main peer (MP) is responsible for the management of the correction process, and the worker peer (WP) performs the correction itself.

As depicted in figure 2, the main peer located in computer X propagates a request on the P2P network asking for the location of all the weather stations (Step 1). This location is represented by the Universal Transverse Mercator (UTM) coordinate system. With the list of weather station returned by the peers (Step 2), the main peer calculates the distance between its location and the locations of all stations and searches for the closest ones. With the closest weather stations list created, the worker peer located in computer X starts searching for gaps in the local climate data. The days in which there are gaps are passed for each the workers peers present in the list generated by the main peer (Step 3)

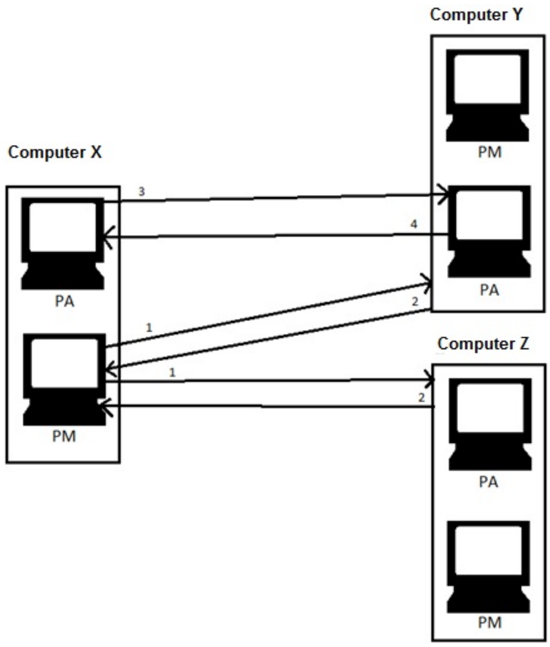

Figure 2 – Correction model

The worker peers of the other stations seek the local data for climatic information from the requested days. For each day, it is done a weighted average of five values, showed in equation 1: the climatic data of the day searched, with weight 2, and the data measured two days before and two days after. The result of this weighted average is returned to the peer that requested the correction (Step 4).

$$local\ average = \frac{x_{-2}+x_{-1}+(x_0*2)+x_1+x_2}{6} \qquad \text{Equation 1}$$

Once all the average values returned by the stations cited in the list created by the main peer are available, the worker peer performs an arithmetic mean, showed in equation 2, where **n** indicates the number of stations and **xn** indicates the result of the weighted average returned by each peer. The result of equation 2 is used for filling the gap found in the respective day.

$$final\ average = \frac{\sum_{n=1}^{x} xn}{n} \qquad \text{Equation 2}$$

For the correction of the artificially created gaps showed in Table 2, two tests were performed: one using climate information from three stations located respectively in *Castro*, *Tibagi* and *Fernandes Pinheiro*, and another using all stations. The results obtained by the correction system were compared to the real climate data and to data generated by stochastic methods that also can be used for filling gaps. This comparison was made by means of Fisher's F and Student´s T tests.

3. RESULTS

The Tables 3, 4, 5 e 6 show the results obtained for the first experiment, which involves three cities providing climatic data for the months of February, May, August and November. All the 6 periods cited in Table 2 are considered. The tables show the values of the average temperature in the entire period for the three values considered: real, stochastic and P2P system generated. It is also showed the value of standard deviation (SD) and the results obtained by the utilization of F and T tests for comparing the two generated values with the real value.

Table 3 – Results of February with 3 weather stations

	Period: 1 - 10			**Period: 11 - 20**			**Period: 19 - 28**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	20,9000	21,4200	21,4400	20,3600	21,0800	20,9944	21,2500	20,4000	21,5306
SD	0,4690	1,1989	0,1646	1,0977	1,5711	0,5916	1,1058	2,1782	0,6366
F test	-	0,0100	0,0046	-	0,3003	0,0797	-	0,0560	0,1155
t test	-	0,2177	0,0054	-	0,2503	0,1250	-	0,2857	0,4957

	Period: 1 - 20			**Period: 9 - 28**			**Period: 1 - 28**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	20,6300	21,2500	21,2172	20,8700	20,8100	21,2661	20,8786	20,9571	21,3026
SD	0,8670	1,3713	0,4805	1,1379	1,8987	0,6582	0,9908	1,7006	0,5604
F test	-	0,0524	0,0134	-	0,0310	0,0214	-	0,0066	0,0042
t test	-	0,0956	0,0117	-	0,9042	0,1858	-	0,8737	0,0552

Table 4 – Results of May with 3 weather stations

	Period: 1 - 10			**Period: 11 - 20**			**Period: 22 - 31**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	16,1800	14,3000	15,8011	14,3600	15,8500	14,6094	14,3300	15,6100	13,6422
SD	2,5059	2,8000	1,5415	2,9740	3,5837	2,2940	3,0793	1,8187	2,4064
F test	-	0,7463	0,1639	-	0,5874	0,4511	-	0,1326	0,4740

t test	-	0,1310	0,6886	-	0,3251	0,8360	-	0,2726	0,5847

	Period: 1 - 20			**Period: 12 - 31**			**Period: 1 - 31**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	15,2700	15,0750	15,2053	14,1650	15,8400	13,9436	14,9548	15,2839	14,6720
SD	2,8348	3,2294	1,9980	2,7848	2,7398	2,1631	2,8514	2,7731	2,1939
F test	-	0,5755	0,1362	-	0,9442	0,2796	-	0,8798	0,1568
t test	-	0,8403	0,9339	-	0,0627	0,7804	-	0,6468	0,6632

Table 5 – Results of August with 3 weather stations

	Period: 1 - 10			**Period: 11 - 20**			**Period: 22 - 31**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	12,9800	17,1900	13,0078	16,4300	15,0700	15,7294	16,0800	14,5900	15,6122
SD	5,5196	4,1162	3,4783	2,6891	5,7579	1,8520	4,8803	6,1578	3,2660
F test	-	0,3952	0,1851	-	0,0332	0,2818	-	0,4993	0,2472
t test	-	0,0691	0,9894	-	0,5072	0,5061	-	0,5562	0,8040

	Period: 1 - 20			**Period: 12 - 31**			**Period: 1 - 31**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	14,7050	16,1300	14,3686	15,8750	14,1250	15,3881	14,8871	15,4806	14,6272
SD	4,5814	4,9913	3,0504	4,3899	5,2755	2,8843	4,8191	5,3155	3,1922
F test	-	0,7126	0,0842	-	0,4305	0,0748	-	0,5947	0,0273
t test	-	0,3528	0,7861	-	0,2613	0,6808	-	0,6467	0,8032

Table 6 – Results of November with 3 weather stations

	Period: 1 - 10			**Period: 11 - 20**			**Period: 21 - 30**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	18,5600	18,5300	18,6572	16,8100	18,9800	18,0044	20,8700	18,7300	20,5889
SD	3,8768	3,1401	1,7055	1,8478	2,5015	0,9299	1,4499	1,6547	0,5896
F test	-	0,5400	0,0225	-	0,3803	0,0531	-	0,7004	0,0132
t test	-	0,9850	0,9429	-	0,0406	0,0845	-	0,0065	0,5771

	Period: 1 - 20			**Period: 11 - 30**			**Period: 1 - 30**		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	17,6850	18,7550	18,3308	18,8400	18,8550	19,2967	18,7467	18,7467	19,0835
SD	3,0891	2,7727	1,3782	2,6365	2,0682	1,5271	3,0391	2,4263	1,5889
F test	-	0,6424	0,0009	-	0,2986	0,0217	-	0,2311	0,0008
t test	-	0,2562	0,4009	-	0,9841	0,5067	-	1,0000	0,5933

By analyzing the results obtained, it can be observed that with three locations the P2P system showed better results in the months of May, August and November when compared to the stochastic generated values.

The values of the F and t tests show that the P2P model could correctly fill the gaps, as the tests show that the real values and the generated values belong to the same population and reflect the same information about the temperature.

Tables 7, 8, 9 e 10 show the results obtained for the second experiment, which involves all cities providing climatic information.

Table 7 – Results of February with all weather stations

	Period: 1 - 10			Period: 11 - 20			Period: 19 - 28		
	Real	Stochastic	P2P	Real	Stochastic	P2P	Real	Stochastic	P2P
Average	20,9000	21,4200	21,4872	20,3600	21,0800	21,2081	21,2500	20,4000	21,8072
SD	0,4690	1,1989	0,1320	1,0977	1,5711	0,5436	1,1058	2,1782	0,6169
F test	-	0,0100	0,0008	-	0,3003	0,0481	-	0,0560	0,0971
t test	-	0,2177	0,0032	-	0,2503	0,0420	-	0,2857	0,1810

	Period: 1 - 20			Period: 9 - 28			Period: 1 - 28		
	Real	Stochastic	P2P	Real	Stochastic	P2P	Real	Stochastic	P2P
Average	20,6300	21,2500	21,3476	20,8700	20,8100	21,4747	20,8786	20,9571	21,4717
SD	0,8670	1,3713	0,4108	1,1379	1,8987	0,6279	0,9908	1,7006	0,5315
F test	-	0,0524	0,0021	-	0,0310	0,0128	-	0,0066	0,0019
t test	-	0,0956	0,0024	-	0,9042	0,0442	-	0,8337	0,0079

Table 8 – Results of May with all weather stations

	Period: 1 - 10			Period: 11 - 20			Period: 22 - 31		
	Real	Stochastic	P2P	Real	Stochastic	P2P	Real	Stochastic	P2P
Average	16,1800	14,3000	15,9508	14,3600	15,8500	14,7722	14,3300	15,6100	13,8853
SD	2,5059	2,8000	1,6244	2,9740	3,5837	2,2629	3,0793	1,8187	2,3302
F test	-	0,7463	0,2125	-	0,5874	0,4279	-	0,1326	0,4189
t test	-	0,1310	0,8110	-	0,3251	0,7313	-	0,2726	0,7200

	Period: 1 - 20			Period: 12 - 31			Period: 1 - 31		
	Real	Stochastic	P2P	Real	Stochastic	P2P	Real	Stochastic	P2P
Average	15,2700	15,0750	15,3615	14,1650	15,8400	14,1463	14,9548	15,2839	14,8533
SD	2,8348	3,2294	2,0102	2,7848	2,7398	2,1101	2,8514	2,7731	2,1635
F test	-	0,5755	0,1430	-	0,9442	0,2356	-	0,8798	0,1361
t test	-	0,8403	0,9069	-	0,0627	0,9810	-	0,6468	0,8750

Table 9 – Results of August with all weather stations

	Period: 1 - 10			Period: 11 - 20			Period: 22 - 31		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	12,9800	17,1900	13,1750	16,4300	15,0700	15,9317	16,0800	14,5900	15,9511
SD	5,5196	4,1162	3,5668	2,6891	5,7579	1,7717	4,8803	6,1578	3,1222
F test	-	0,3952	0,2094	-	0,0332	0,2298	-	0,4993	0,1994
t test	-	0,0691	0,9263	-	0,5072	0,6305	-	0,5562	0,9447

	Period: 1 - 20			Period: 12 - 31			Period: 1 - 31		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	14,7050	16,1300	14,5533	15,8750	14,1250	15,6788	14,8871	15,4806	14,8723
SD	4,5814	4,9913	3,0843	4,3899	5,2755	2,7583	4,8191	5,3155	3,1706
F test	-	0,7126	0,0928	-	0,4305	0,0494	-	0,5947	0,0249
t test	-	0,3528	0,9029	-	0,2613	0,8665	-	0,6467	0,9887

Table 10 – Results of November with all weather stations

	Period: 1 - 10			Period: 11 - 20			Period: 21 - 30		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	18,5600	18,5300	18,4722	16,8100	18,9800	18,0739	20,8700	18,7300	20,3503
SD	3,8768	3,1401	1,1872	1,8478	2,5015	0,8411	1,4499	1,6547	0,5637
F test	-	0,5400	0,0016	-	0,3803	0,0282	-	0,7004	0,0096
t test	-	0,9850	0,9467	-	0,0406	0,0646	-	0,0065	0,3122

	Period: 1 - 20			Period: 11 - 30			Period: 1 - 30		
	Real	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**	**Real**	**Stochastic**	**P2P**
Average	17,6850	18,7550	18,2731	18,8400	18,8550	19,2121	18,7467	18,7467	18,9655
SD	3,0891	2,7727	1,0220	2,6365	2,0682	1,3599	3,0391	2,4263	1,3322
F test	-	0,6424	0,0000	-	0,2986	0,0059	-	0,2311	0,0000
t test	-	0,2562	0,4272	-	0,9841	0,5792	-	1,0000	0,7199

The results obtained by the correction model with all stations involved are similar to the results obtained for three stations, with the exception of the period of the 10 last days of February.

It can be noted that the F and t tests also showed that the P2P model generated climatic data that can be considered to be statistically contained in the same population.

4. CONCLUSION

It can be concluded that the use of the proposed correction model allowed more accurate results, when compared to stochastic generated data, for the months contained in autumn, winter and spring seasons.

It was also observed that the amount of weather stations considered can influence the results, although with three stations it was already possible to achieve satisfactory results.

This paper also showed that the proposed model was able to generate climate data that can statistically be considered as belonging to the same population of the real data, so it can be safely used for planning agricultural activities.

ACKNOWLEDGEMENTS

The authors want to thank SIMEPAR and ABC Foundation for providing the data used in this research.

REFERENCES

[1] HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. Agricultural and Forest Meteorology, v.103, p.137-157, 2000.

[2] MATEO, M. A. F.; LEUNG, C. K. S. Design and Development of a Prototype System for Detecting Abnormal Weather Observations. Proceedings of the 2008 C3S2E conference, 2008.

[3] MOON. J.; CHO Y. A Point-Based Inventive System to Prevent Free-Riding on P2P Network Enviroments. Computational Science and its Applications – ICCSA, p.462-471, 2011.

[4] SENGER, L. J.; SOUZA, M. A.; FOLTRAN, D. C. Towards a peer-to-peer framework for parallel and distributed computing. 22nd International Symposium on Computer Architecture and High Performance Computing. p.127-134, 2010.

[5] SENTELHAS, P. C; MONTEIRO, J. E. B. de A. Agrometeorologia do Cultivos. Brasília, DF: INMET, 2009.

[6] STEINNHAEUSER, K; CHAWLA, N. V; GANGULY, A. R. An Exploration of Climate Data Using Complex Networks. ACM SIGKDD Exploration Newsletter. v.12, p.25-32, 2010.

[7] TARDIVO, G. ; BERTI, A. A Dynamic Method for Gap Filling in Daily Temperature Datasets. *J. Appl. Meteor. Climatol.*, v. 51, n. 6, p. 1079–1086, 2012.

[8] TARDIVO, G. ; BERTI, A. The selection of predictors in a regression-based method for gap filling in daily temperature datasets. *International Journal of Climatology*. v.34, n. 4, p. 1311-1317, 2014.

[9] ZANETTI, S. S.; OLIVEIRA, V. P. S.; PRUSKI, F. F. Validação do modelo ClimaBr em relação ao número de dias chuvosos e à precipitação total diária. Engenharia Agricola, v. 26, p.96-102, 2006.