Optimization of an urban rain gauge network utilizing different PCA-Variable identification techniques

H41L-2239

Juan P. Nogués[1*], Eduardo J. Rodríguez[2], Maria A. Bellassai[3], Hugo A. Falcon[4]

AGU Fall Meeting, Washington DC, December 2018

Abstract

The Metropolitan Area of Asunción (AMA) that consists of 11 cities, and an urban population of 2.3 million people over 640 km^2, only has a storm drain collection network that covers 4% of its area. This lack of coverage, adding to its poor land use planning, creates havoc during normal to intense rainfall events producing millions in damages.

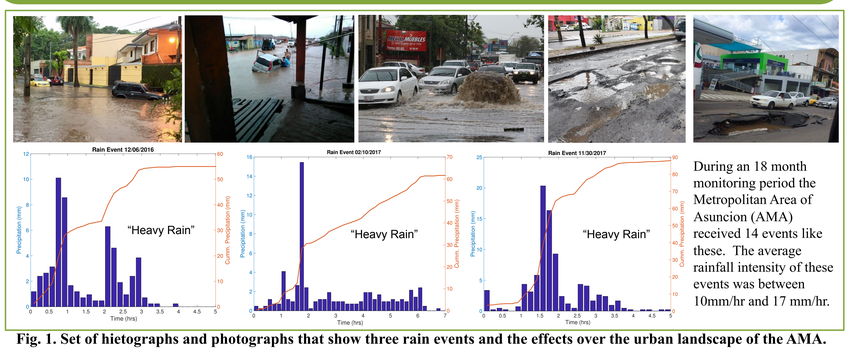
In order to study the effects of urban runoff, through numerical models and the design of waterworks, appropriate rainfall information needs to be gathered and analyzed. With this purpose a 27 rain gauge network has been deployed over the AMA.

Here we present a principal component analysis (PCA) on daily rainfall data collected from June 2016 to December 2017 in order identify the stations that should be kept in place, given the objective of accurately capturing the spatial variability of rainfall over the Metropolitan Area of Asuncion.

Missing data, which amounted to around 17% of all the collected data, was filled by applying an inverse weighted distance method. Outliers were identified and eliminated utilizing Grubbs test. The PCA was done utilizing dates with rain events that on average recorded more than 10 mm of rainfall. In order to determine the most important components, the Kaiser test was performed which specifies the amount of components to keep (i.e. those with eigenvalues above or equal to 1).

In order to pair the principal components to real variables (rain gauge stations) different variable identification methods were tested. After selecting the set of stations, the original data set was interpolated using those selected stations. The root mean square, the coefficient of determination and the Nash-Sutcliffe coefficients were calculated as evaluation methods. Of the 6 methods used, the 'Correlation between X and Z' method (Method 2) and the 'Varimax' method (Method 4) seem to be the most adequate at representing the mean values, and having the highest r^2 values and the lowest RMSE values. The 'Covariance' method (Method 1) and the 'Loading Combination' method (Method 3) performed the worst.

Problem Statement



During an 18 month monitoring period the Metropolitan Area of Asuncion (AMA) received 14 events like these. The average rainfall intensity of these events was between 10mm/hr and 17 mm/hr.

Fig. 1. Set of hietographs and photographs that show three rain events and the effects over the urban landscape of the AMA.

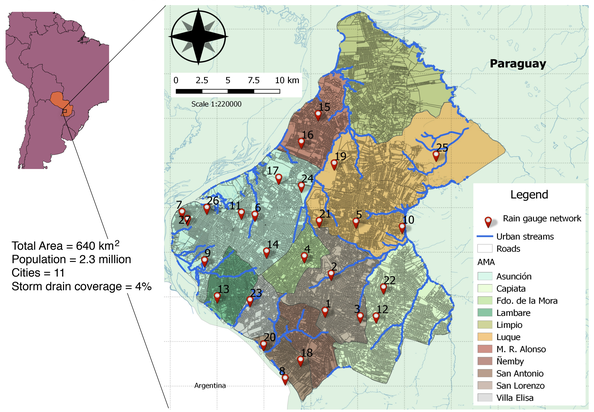


Fig. 2. Map of study area with the 27 rain gauge stations showing. Each station has two tipping bucket rain gauges that have collected data for 18 months, between 2016 -2017.

Selection, of how many and which rain gauge stations to keep is an important question to answer when considering the value of the spatial-temporal variation of the rainfall and the costs of operating each station. By definition both objectives work against each other − one would want to maximize former while reducing the latter.

Fig. 2 shows the map of the current rain gauge **network with 27 stations**. The rainfall data gathered through can be utilized to design water works the city desperately needs as wells a providing other valuable information to calibrate doppler radars, perform water budgets, and calibrate hydrological and atmospheric numerical models.

Station density is 4 stations for every 100 km^2 and the average distance between them is of 12 km^2.

Methods

Data Selection: The rainfall data collected over the 18 months had missing days because either the stations had not been equipped with the gauges, or there was a systematic error with the data that made it doubtful (i.e. clogs, tilted gauges, bird nests, etc.) − see Fig. 3. Of the 18 months of recorded data, 350 days had a rainfall event. Not all gauges recorded an event, indicating the variability in space of the study area. In order to perform a systematic data analysis, maximum outliers were also removed by performing **Grubb's test**. This test eliminated a total of 10 points.

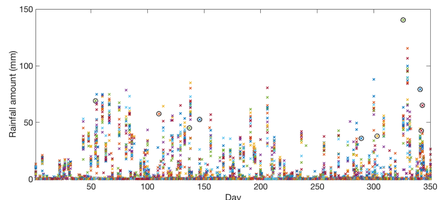


Fig. 3 Record of daily rainfall events at each station. Each cross represents a rainfall event measured at a station. Circled points are outliers identified through Grubb's test, that were eliminated.

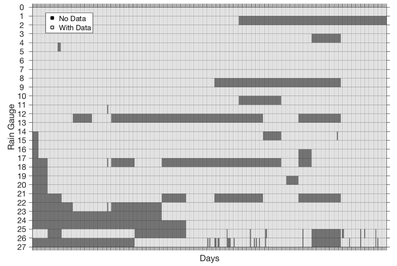
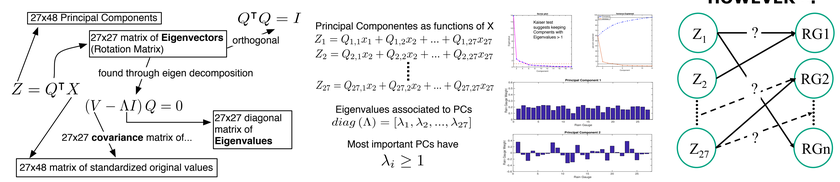


Fig. 4. List of days with data, and no data, of all 27 stations for the 350 days that recorded a rain event.

Data Filling: Once the data was cleared of outliers, the stations with missing data were filled utilizing **the inverse distance weighted** (IDW) method. This interpolation method method has been shown to be as good or better, than Ordinary Krigging, in error reduction when dealing with high density rain gauge networks (Keblouti et al., 2012).

Data Reduction: Given that the proposal was to work utilizing principal components which is a variance reduction method, only rain data of important events were kept in order to reduce the influence of low variance events. This was done by **eliminating days/ events that had an average less than 10 mm of rain**. This procedure eliminated 305 days (87%) of the data and left a **set of 48 rainfall events** where all stations had either a recorded or interpolated information.

Principal Component Analysis (PCA): The objective of PCA is to **simply the description of the original data set** by retaining only the most important variables (Manly, 2004). The basic process involves calculating the eigenvectors and eigenvalues of **the covariance matrix** of the standardized original data (i.e. zero mean and unit variance).



After doing all the algebraic manipulations, we are left with a matrix of component weights that relate the original values to the "new" principal components. **However it is not clear, or obvious, how to translate the principal components to real variables.** There are several methods to this (Al-Kandari & Jolliffe, 2001). We propose to study a combination of methods:

Method 1 − Largest Variance

VC Method - Keep the **k** variables that have the highest variances in the sample covariance matrix.

The **PCA** gives only the number of variables to keep (**k**) through the study of the eigenvalues (Kaiser Test).

Method 2 − Correlation between X and Z

XvZ Method − Correlate the first **k** PCs (**Z**) with the original values (**X**). Grab the **k** variables with the highest absolute correlations.

$$G = Corr\,(Z, X)$$

Method 3 − Loading Combination

NLC Method − Grab the first **k** PCs in **Q**, take the average of the absolute values of each variable and keep the k variables with the highest value.

The **PCA** gives the number of variables to keep (**k**), through the study of the eigenvalues (Kaiser Test), and the matrix **Q**.

Method 4 − Varimax Rotation

Varimax Method− Most common approach, consists of rotating the **Q** matrix, so that it keeps its orthogonality condition.

Varimax searches for a linear combination of the factors in Q such that the variance of the squared loading is maximized.

Method 5 – Summation

Count up all the times the variables selected using methods 1 through 4 were selected. Keep the **k** variables that have appeared the most.

Method 6 – Ranking of Selection

Rank each variable selected in the methods 1 through 4 with values from 1 to **k**. The **k** weight/rank corresponds to the most "important" variable (given by the 1st PC). Sum all the ranks/ weights and select the **k** variables with the highest ranks.

Results

Evaluation of the Methods − In order to evaluate the efficiency of the method the original 48 day data was recreated with **k** retained variables using the IDW method. **For our study k=5** (i.e. 5 gauges are retained).

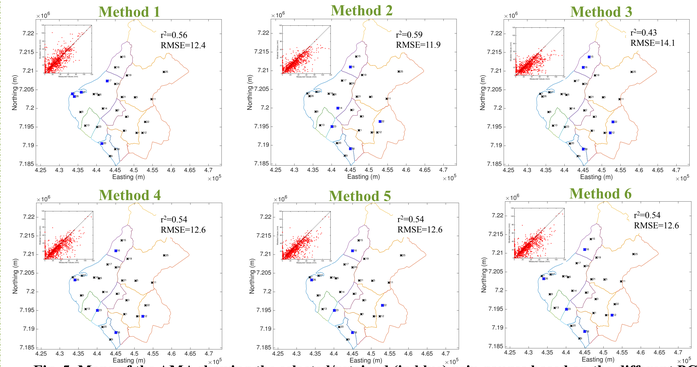


Fig. 5. Maps of the AMA showing the selected/retained (in blue) rain gauges based on the different PCA-variable selection methods. Also shown are the scatter plots between observed and modeled data (using IDW interpolation) with the remaining wells. Scatter plots also show the r^2 and RMSE values.

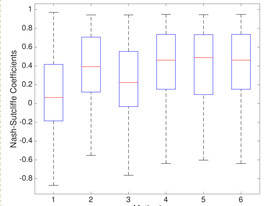


Fig. 6. Boxplot of the Nash-Sutcliffe coefficients produced by comparing the observed data with the modeled data using 5 stations.

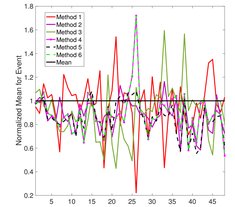


Fig. 7. Normalized mean values for the 48 rain events using only the 5 retained station for each method. Values were compared to original mean.

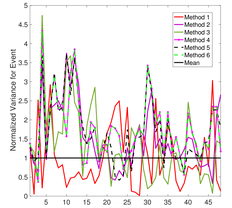


Fig. 8. Normalized variance values for the 48 rain events using only the 5 retained stations for each method. Values were compared to original variance.

Conclusions

- All the proposed methods seem to perform similarly, which is mainly due to the fact that there is a very high correlation among all stations. If different parameters (not only total rainfall), with lower correlations among them, was considered the methods might differ more.
- Of all the methods, method 1 and 3 perform the worst, most likely because they fail to capture correlations among the stations. Method 2 (Correlation) and Method 4 (Varimax) perform the best.
- Station 16, 23 and 26 seem to be the most important since they appear in the majority of the methods.
- Method 5 and Method 6 are almost identical, they only differ in one station. If more methods were to be used their difference would become more apparent.

Authors

1 Universidad Paraguayo Alemana, San Lorenzo Paraguay * Corresponding author, **juan.nogues@upa.edu.py**
2 Undergraduate student in the Industrial Engineering program at the Universidad Paraguayo Alemana
3 Undergraduate student in the Environmental Engineering program at the Universidad Nacional de Asunción
4 Undergraduate student in the Environmental Engineering program at the Universidad Catolica de Asunción

ACKNOWLEDGEMENTS: This research is funded by PROCIENCIA initiative of the CONACYT-Paraguay under Project 14-INV-189

References

Keblouti, M., Ouerdachi, L., and Boutaghane, H. (2012). Spatial interpolation of annual precipitation in annaba-algeria-comparison and evaluation of methods. Energy Procedia, 18, 468–475.

Manly, B. (2004). Multivariate Statistical Methods: A Primer. London. Chapman & Hall/CRC

Al-Kandari N. M. and I. T. Jolliffe (20019). Variable Selection and Interpretation of Covariance Principal Components. Communications in Statistics − Simulation and Computation, 30:2, 339-354.