Improved Homogeneous Group Formation for Regional IDF Estimation

Zhe (Emma) Yang and Donald H. Burn

Department of Civil & Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada

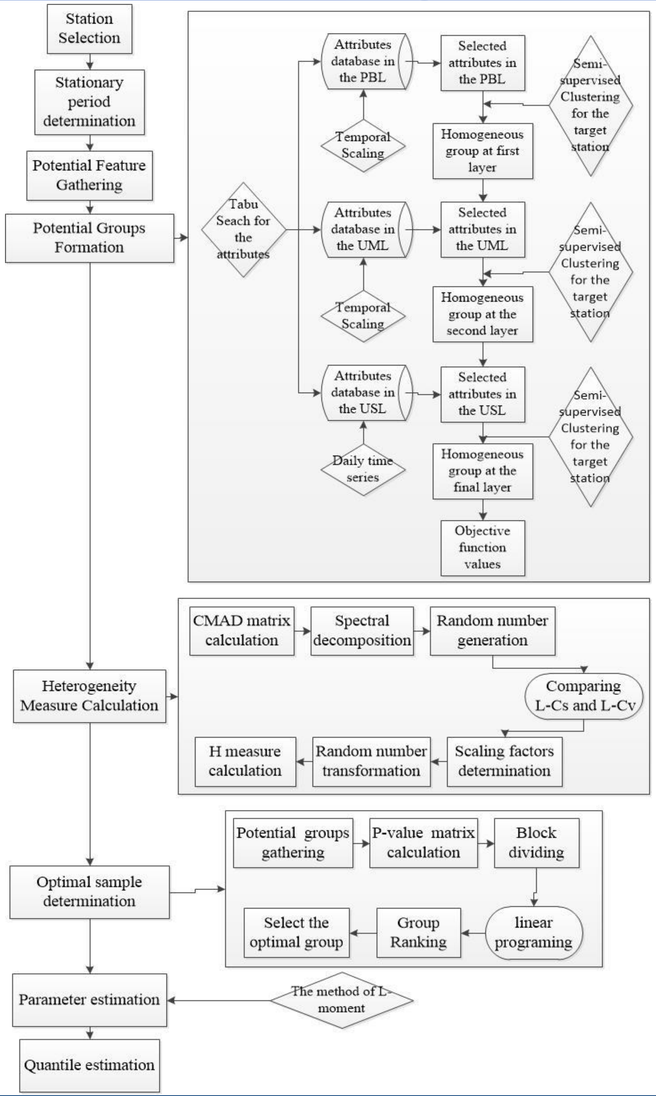
Abstract

Extreme rainfall events can have devastating impacts on society. However, recent changes in rainfall climatology caused by climate change and urbanization have made estimates provided by the traditional IDF approach become increasingly inaccurate. Three major problems exist in the traditional IDF estimation: the ineffective choice of attributes in the formation of a homogeneous group, an inadequate number of stations in the pooling group for quantile estimation and the negative impacts from pooling group's cross-correlation on the homogeneity test. For the first issue, an automatic feature selection and weighting algorithm, specifically the hybrid searching algorithm of Tabu search and supervised clustering, was used to select the relevant features for homogeneous group formation at a specific region. During the process, the impacts of urbanization and climate change on rainfall climatology were considered. For the second issue, the two sample Kolmogorov-Smirnov test-based sample ranking process is used to compare the confidence interval widths generated from the potential groups, during which the method of linear programming is used to rank these groups. To generate the cross-correlated random number for the last issue, the mean absolute difference matrix is used in the Eigen decomposition to obtain the relationship among the input stations and the Gaussian random datasets from this representation are transformed into a non-Gaussian distribution. The comparison of L- skewness and L- kurtosis between the generated groups and original groups is used as the performance indicator to scale the generated series.

Objectives

- Effective homogeneous group formation
- Optimal pooling group selection
- Heterogeneity measurement improvement

Framework of the Algorithm



Methodology

Automatic Feature Selection and Weighting in Formation of Homogeneous Group

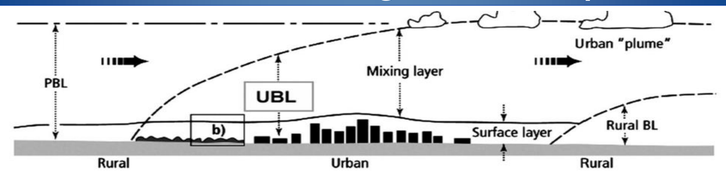


Figure 1: the vertical structure of the Planetary Boundary Layer

Procedures:

1) Original Feature Gathering. The potential rainfall-related feature values at each layer with different resolutions are extracted at appropriate temporal steps for stations in initial group.
2) Feature selection. Based on memory recorded in the search algorithm, the optimal feature combinations at each layer are used as similarity indicators in the pooling at each layer.
3) Feature weighting. The selected features are weighted through the method of Lagrange multiplier before being used in the formation of the homogeneous group.
4) Semi-supervised Clustering. To generate the appropriate pooling group for the target station, fuzzy c-means clustering is conducted with two extra constraints at each layer.
5) Homogeneous Region Formation. The weighted optimal features are used to generate ideal pooling group for the target station through step (4). The pooling group generated from a higher layer is used as the initial input group for the clustering in a lower layer.

Heterogeneity measurement improvement

To account the cross-correlation impact on the heterogeneity measurement calculation, new method is proposed to generated cross-correlated random number for the H test:

1) Cross mean absolute difference (CMAD) matrix calculation. Instead of the traditional cross-correlation matrix, the cross mean absolute difference matrix is used to describe the cross relationship among the input stations. $\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j||y_i - y_j|$ $\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j||y_i - y_j|$

$$r_m = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j||y_i - y_j|}{\left(\sum_{i}^{n}\sum_{j}^{n}|x_i - x_j|\right)\left(\sum_{i}^{n}\sum_{j}^{n}|y_i - y_j|\right)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j||y_i - y_j|}{(2L - scale_x)(2L - scale_y)}$$

2) Spectral decomposition. Based on the matrix C obtained from step (1), Eigen decomposition is conducted to obtain Eigen values and the corresponding vectors.

$$\widehat{\mathbf{C}} = \mathbf{\Phi}_1^{\mathbf{C}} \mathbf{\Lambda}_1^{\mathbf{C}} [\mathbf{\Phi}_1^{\mathbf{C}}]^{\mathrm{T}}$$

3) Cross- correlated random number generation. Multiply the random Gaussian number with the above Eigen values and the corresponding vectors to obtain the Cross- correlated Gaussian random number:

$$\boldsymbol{\chi}^{\mathbf{D}} = \mathbf{\Phi}^{\mathbf{D}} (\mathbf{\Lambda}^{\mathbf{D}})^{1/2} \boldsymbol{\xi}$$

4) Scaling factors determination. Determine the multiply factors for the mean and standard deviation to scale the generated random number from step (3). During the process, the comparison of L- skewness and L- kurtosis between the generated groups and original groups is used as the performance indicator to select the appropriate multiply factors .
5) Random number transformation. Transform the obtained random number to the target distribution.

Optimal homogeneous group selection

Two Sample Kolmogorov-Smirnov Test is used to determine if two samples come from the same distribution:

$$H_0 : F_1 = F_2 ; \quad H_1 : F_1 \neq F_2$$

The Kolmogorov–Smirnov statistic is:

$$D_{n,m} = \max_x |F_{1,n}(x) - F_{2,m}(x)|$$

When $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$ or $p-value < 0.05$, the null hypothesis is rejected at level α.

The p-value can reflect the extent of similarity between the input two samples. Thus the p-value is used as the distribution similarity indictor in the following linear programing for the selection of optimal homogeneous group. The procedures are described as the following:

1) Potential optimal groups gathering. Based on the similarity measurement provided by the previous procedure, the potential optimal groups at different sizes are selected.
2) P-value matrix calculation. Based on the two sample KS test, the p-values between two of the potential groups are calculated and used to form the p-value matrix.
3) Block dividing. Separate the groups whose p-values are close to 1, then consider these groups as new one and substitute their original p-values with average p-values.
4) Ranking the groups obtained from step (3). Following the rule of "Shortest path problem" to find the path to reach the largest sum of p-values. During the process, each group can only appear in the path once.

Case Study

❖ Study area

This application is conducted based on the 82 IDF stations in Ontario and Quebec. The target station is the City of Toronto.

❖ Dataset

NOAA Global Ensemble Forecast System Reforecast (GEFS/R) and ERA-Interim Database.

The target layer	Potential features
The PBL (Planetary Boundary Layer)	Air temperature, Geopotential height, Specific humidity, U-component and V- component of the wind velocity (at the 300hPa, 500hPa and 700hPa pressure level), and Convective Available Potential Energy (CAPE)
The UML (Urban Mixed Layer)	Air temperature, Geopotential height, Specific humidity, U-component and V- component of the wind velocity (at the 850hPa and 925hPa pressure level), and Vertical integral of water vapor (VIWV).
The USL (Urban Surface Layer)	Urban surface sensible heat flux (SHTFL), Urban surface latent heat flux (LHTFL), Photosynthetically Active Radiation index (PAR), Surface Net Solar Radiation (SNSR), Surface Net Thermal Radiation (SNTR) and the Surface Roughness (SR).

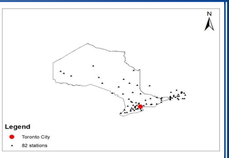


Figure 2. Location map showing the 82 IDF sites

❖ Automatic Feature Selection and Weighting in Formation of Homogeneous Group results

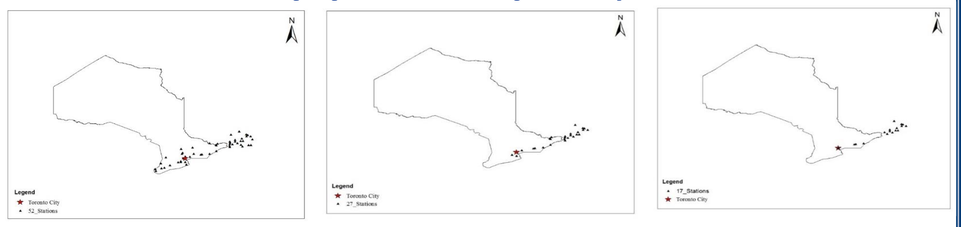


Figure 3 Graphic display of homogeneous group of 52, 27 and 17 weather stations obtained from the clustering different layers

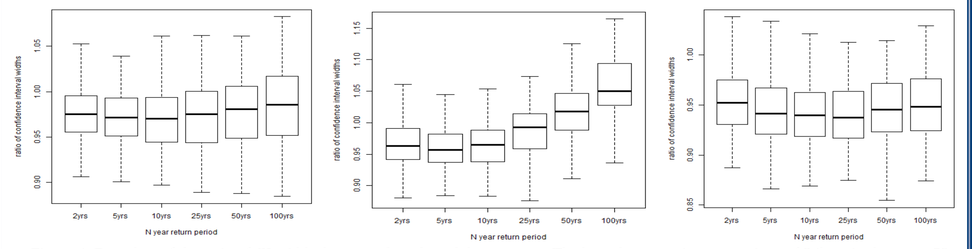


Figure 4 Box plots of the ratio of CI widths between the selected three groups. The box plots contains comparison of 24h series between 52 stations over 82 stations, 27 stations over 52 stations, 17 stations over 27 station.

❖ Heterogeneity measurement

The potential homogenous group are the ones with the sample size of 17, 27, 37, 39, 52 and 82. Since the proposed method can only applicable in the even series, equal rainfall series have been selected.

Groups	17	27	37	39	52	82
Traditional H_1 test	0.41	-0.56	-0.49	-0.91	-0.79	-0.56
Proposed H_1 test	0.80	0.13	0.34	0.21	0.43	0.36

❖ Optimal homogeneous group selection

P-value	17	27	37	39	52	82
17	1	0.892	0.7653	0.7499	0.2453	0.1355
27	0.892	1	0.9993	0.9997	0.7435	0.3407
37	0.7653	0.9993	1	1	0.9913	0.7904
39	0.7499	0.9997	1	1	0.9441	0.553
52	0.2453	0.7435	0.9913	0.9441	1	0.9814
82	0.1355	0.3407	0.7904	0.553	0.9814	1

P-value	17	Group	52	82
17	1	0.8024	0.2453	0.1355
Group	0.8024	1	0.892967	0.561367
52	0.2453	0.892967	1	0.9814
82	0.1355	0.561367	0.9814	1

Group: the average p-value of group 27,37 and 39

Based the second p-value matrix, the width of confidence interval (CF) of above groups can be ranked as the following from small to high :

CF(17) < CF(Group) < CF(52) < CF(82)

CF widths comparison within the Group can be conducted by using either one of the following rules:

1) Among the groups that share similar distributions, the ones with large sample size tend to have narrower CF widths than the ones with smaller sample size.
2) Conduct one-side two sample KS test to determine CDF of the group that lies above the rest, and the higher CDF usually generate narrower CF width.

Thus the ranking can be: CF(17) < CF(39) < CF(37) < CF(27) < CF(52) < CF(82)

Conclusions & Future work

The proposed algorithm will be applied in the nonstationary environment in the following research

Contact Information

Zhe (Emma) Yang, PhD Candidate
Department of Civil & Environmental Engineering
University of Waterloo, Waterloo, Ontario, Canada
Tel: (519) 721-9755
Email: z232yang@uwaterloo.ca