The 2014 World Congress on
Advances in Civil, Environmental, and Materials Research (ACEM14)
Busan, Korea, August 24-28, 2014

Mixture Probability Distributions of Wind Pressure on Low-rise Buildings

*Haitao Zheng[1)] Guoqing Huang[2)] and Xiaobo Liu[3)]

[1),3)]*Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu 610031, China*
[2)] *School of Civil Engineering, Southwest Jiaotong University, Chengdu 610031, China*
[1)] *htzheng@gmail.com*

ABSTRACT

Two types of mixture extreme distributions are adapted to fit the wind pressure coefficients over the roof surface of low-rise buildings. Parameters of the mixture distribution are estimated by using Maximum likelihood. Two taps over the roof surface of low-rise building are selected for the data analysis. Fittings of the tail part of the data are compared via several measurements. Data analysis for two taps shows that the parametric hybrid GPD distribution generally may not be better than the two components mixture distributions and EVD_lognormal mixture distribution may be a better choice to describe the wind pressure distribution over the roof surface of low-rise buildings in this study.

Keywords: Wind pressure coefficient; hybrid GPD; mixture distribution; goodness of fit; PPCC;

1 INTRODUCTION

In USA and other areas around the world, many low-rise buildings are damaged by extreme wind events and much of wind-induce damage on those low-rise buildings could be attributed to roof failures. Roof damage is related to various wind-related influential factors. One of those factors is wind pressure. Therefore, to better evaluate the wind-induce damages, it is important to model the maxima wind pressure coefficients. Literatures show that the distribution of pressure coefficients is non-Gaussian and various extreme value distributions and mixture distributions are proposed to model the wind pressure.

[1,2)] Professor

[3)] Graduate Student

Extreme value theory is widely used to describe the likelihood of unusual behavior or rare events occurring, such as in financial study, insurance, hydrology or environmental applications, etc. A good reference book that discusses different applications with extreme value theory is by Reiss and Thomas (2001). However, a single distribution cannot fit the data well in many situations. Researchers developed mixture distributions to model the non-Gaussian data. There two types of mixtures: 1. The data comes from two independent distributions; 2. Bulk part of the data is from one distribution and the rest of the data comes from one or two extreme distributions. There are some related literatures on both types. Bordes L, et al (2006) studied the two-component normal mixture and corresponding estimation theory. Benaglia T, et al (2009c) created an R package which incorporates parametric and non-parametric mixture distributions modeling. Akdag et al. (2010) discussed the two-component mixture Weibull distribution to estimate wind speed characteristics. Kollu et al. (2012) studied three different mixture distributions for wind speed.

For the second type, many researchers studied extreme value mixture models so that the mixture models can provide both the estimation of the threshold and parametric and/or non-parametric estimation of the bulk and tail part of the data. The extreme value mixture models typically have two components:

1 a parametric/nonparametric model for describing all the non-extreme data below the threshold (that part of the data are referred to as bulk part);

2 a typical extreme value model for modeling data above the threshold (the data points over the threshold are referred to as the tail part).

Carreau and Bengio (2008) proposed to use mixture model (referred to as Hybrid GPD) with any distribution for the bulk part and GPD for the tail part and the model just needs a continuity constraint at the threshold. MacDonald et al. (2011) and MacDonald et al. (2013) developed a non-parametric kernel density estimator based extreme value mixture models extending on that developed by Tancredi et al.(2006). Hu, Y. (2013) developed an R package that analyze data from a single population with parametric/nonparametric mixture distributions.

In the study, we mainly compare the hybrid GPD to the mixture distributions for the tail part of the data. The process of the analysis is as follow. We use Normal hybrid-GPD to fit the pressure data to get threshold and the GPD of the tail part of the data. The fitted hybrid PGD serves as a benchmark. Next, various mixture distributions are used to fit the same data, including mixture of normal-GEV, normal-Gamma, GEV-lognormal, Normal-Weibull, EVD-Weibull, etc. After we fit the data with the above distributions, the fitted distributions are used to predict the probabilities and quantiles above the threshold which is estimated from the normal hybrid GPD method. Finally, five methods of goodness-of-fit are used to measure the closeness of the fitted distribution to the empirical one. Note that the measurements are applied for the tail part of the data.

This paper is organized as follows. In section 2, hybrid GPD and various mixture distributions are proposed to model wind pressure coefficient data and 5 measurements of goodness-of-fit are given to evaluate the proposed density functions. In section 3, several representative points are selected and fitted by the proposed distributions for the tail part of the data. Finally, we conclude our findings.

2 MATERIALS AND METHODS

Classical theoretical results are concerned with the stochastic behavior of some maximum (minimum) of a sequence of random variables which are assumed to be independently and identically distributed.

von Mises (1954) and Jenkinson (1955) proposed a way of unifying the three different types of extreme value distributions, which lead to the generalized extreme value distribution *GEV(μ, σ, ξ)*. The distribution function of *GEV(μ, σ, ξ)* is given by:

$$F(x|\mu, \sigma, \xi) = \begin{cases} exp\left\{-[1 + \xi\left(\frac{x-\mu}{\sigma}\right)]\right\}^{-1/\xi} & \xi \neq 0 \\ exp\left[-exp(\frac{x-\mu}{\sigma})\right] & \xi = 0 \end{cases} \qquad (1)$$

The block maxima based the GEV distribution is inefficient as it is wasteful of data when the complete dataset (or at least all extreme values) are available. One way of overcoming these difficulties is to model all the data above some sufficiently high thresholds. Such a model is commonly referred as the peaks over threshold or threshold excess model. The advantage of the threshold excess model is that it can make use of the all the extreme" data. Under certain conditions, Coles (2001) show that the excess $X - u$ over a suitable u can be approximated by the generalized Pareto distribution (GPD) *G(x|u, σ, ξ)*:

$$G(x|u, \sigma_u, \xi) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ 1 - exp\left[\left(\frac{x-u}{\sigma_u}\right)\right]_+ & \xi = 0 \end{cases}, \qquad (2)$$

Where $x > u, \sigma_u > 0, 1 + \xi\left(\frac{x-u}{\sigma_u}\right) > 0$ and σ_u reminds us the dependence between scale parameter σ_u and threshold *u*.

The density function of the hybrid Pareto model with single continuity constraint is therefore defined as:

$$f(x|\mu, \beta, u, \xi) = \begin{cases} \frac{1}{\tau} \hbar(x|\mu, \beta) & x \leq u \\ \frac{1}{\tau} g(x|u, \sigma_u, \xi) & x > u \end{cases}. \qquad (3)$$

The distribution function (CDF) of the hybrid Pareto model with single continuity constraint is defined as:

$$F(x|\mu,\beta,u,\xi)=\begin{cases}\frac{1}{\tau}H(x|\mu,\beta) & x\leq u\\ \frac{1}{\tau}[H(x|\mu,\beta)+G(x|u,\sigma_u,\xi)] & x>u\end{cases}, \quad (4)$$

where $H(.|\mu,\beta)$ is the parametric or nonparametric distribution function and GPD distribution function is defined as $G(x|u,\sigma_u,\xi)$. The τ is the usual normalizing constant and $\tau=1+H(u|\mu,\beta)$, where the 1 comes from the integration of the unscaled GPD.

Another possible method is to use the mixture distribution to model the wind pressure data. Here, we consider two-component mixture distribution:

$$f(x|\theta)=\pi f_1(x|\theta_1)+(1-\pi)f_2(x|\theta_2), \quad (5)$$

where $\theta=(\theta_1,\theta_2)^T$, $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$ are two given probability density functions and $0\leq\pi\leq 1$. Maximum likelihood method can be used to estimate the parameters for the mixture distributions.

Goodness-of-fit tests

Goodness-of-fit tests are used to measure the deviation between the predicted data using theoretical probability function and the observed data. In this paper, five statistical tests are considered as judgment criteria to evaluate the fitness of PDFs (Filliben (1975), Kollu et al. (2012)).

PPCC test:

The PPCC test was developed by Filliben (1975) and is known as a simple and powerful goodness-of-fit test. The test statistic is the correlation coefficient *r* between the ordered observations x_i and the corresponding fitted quantiles m_i determined by proposed probability distribution. If the assumption that the observations could have been drawn from the fitted distribution is true, then the value of *r* is close to 1.0. The correlation coefficient r is given by (Filliben, 1975)

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(m_i-\bar{m})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(m_i-\bar{m})^2}} \quad (6)$$

Kolmogorov-Smirnov test:

The Kolmogorov-Smirnov test (K-S) is defined as the maximum error in cumulative distribution functions (20).

$$\mathrm{KS}=\max|\widehat{F_i}-F_i| \quad (7)$$

where $\widehat{F_i}$ and F_i are the fitted and Empirical cumulative distribution functions, respectively. Lesser K-S value indicates better fitness.

R^2 test:

A larger value of R^2 indicates a better fit of the estimated cumulative probabilities to the empirical cumulative probabilities. R^2 is defined as:

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{F_\iota} - \bar{F})^2}{\sum_{i=1}^{n}(\hat{F} - \bar{F})^2 + \sum_{i=1}^{n}(\widehat{F_\iota} - F_i)^2}, \tag{8}$$

where $\bar{F} = \frac{1}{n}\sum_{i=1}^{n}\widehat{F_\iota}$.

Chi-square error:
$$\chi^2 = \sum_{i=1}^{n}\frac{(\widehat{F_\iota} - F_i)^2}{\widehat{F_\iota}}, \tag{9}$$

Root Mean Squared Error:

Root mean squared error (RMSE) provides a term-by-term comparison of the actual deviation between observed and predicted probabilities. A lower value of RMSE indicates a better distribution function model.

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{F_\iota} - F_i)^2\right]^{1/2} \tag{10}$$

3 DATA ANALYSIS

In this study, the pressure coefficient data is obtained from the website: http://fris2.nist.gov/winddata/uwo-data/uwo-data.html. Before the analysis, the original data are preprocessed. Details can be found in Ho et al. (2003).

The prototype building is located on the suburban terrain with size of 125ft × 80ft × 32ft or 38.1m × 24.4m × 9.8m (Length × Depth × Height) and roof slope of 3:12. The data are collected from the wind tunnel study with 1:100 test model. The sampling frequency is 500 Hz and the wind speed at roof height is set at 110 mph or 49.2m/s. The locations of sensors on the roof can be found in Figure 1.

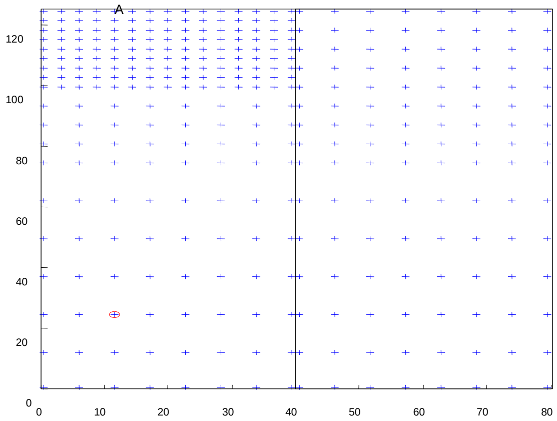

Fig 1: distribution of taps on the roof

The skewness and kurtosis of the wind pressure coefficient at each tap from 13 AOAs (180 deg to 360 deg with the incremental of 15 deg) are plotted in Figure 2. From figure, the non-Gaussian property of pressure coefficient is clear for many taps.

Following the process mentioned previously, the wind pressures at two taps (skewness=3.578, 4 and Kurtosis= 41.95, 32.09) are fitted with hybrid GPD (bulk part is fitted with normal and tail part is fitted with GPD, see formula (4)) and various mixture distributions(see formula (5)). Note that the normal distribution was chosen since the pressure coefficient is not nonnegative. The two component mixture distributions we've chosen are: distribution of Normal and Gamma mixture, distribution of Normal and EVD mixture, distribution of EVD and lognormal mixture, distribution of EVD and Weibull mixture and distribution of Normal and Weibull mixture. For comparison purpose, we also fit the data with normal kernel density method. The fitted kernel density function and density of hybrid GPD are plotted in the following figures. In the analysis, the data are fitted with Hybrid GPD first to get threshold, then the quantils and $\widehat{F_i}$s for each mixture are estimated for all the data points over the threshold and all the measurements of goodness-of-fit are applied to the tail part of the data(right side of the vertical lines in the histograms). The results of goodness-of-fit are summarized in the following tables.

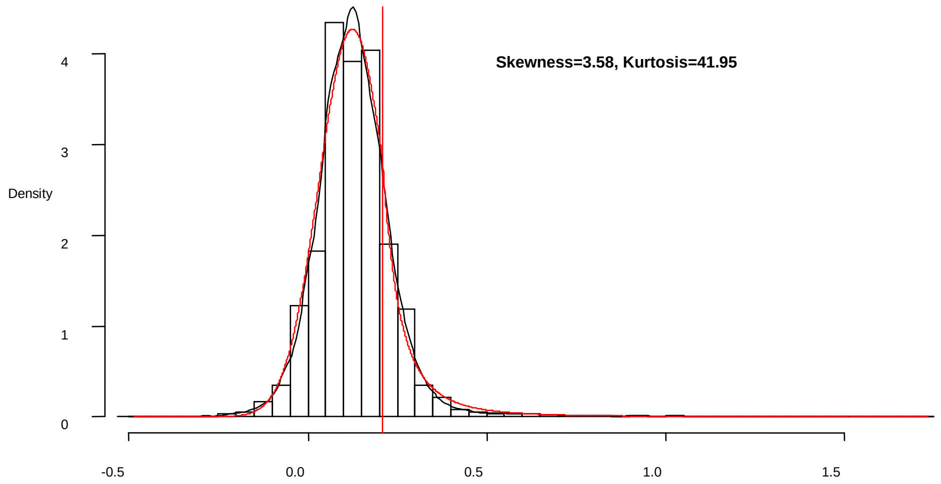

Fig 2: Histogram and fitted curves for Tap 1

	Skewness=3.58, Kurtosis= 41.95				
	PPCC	K-S	R2	Chi2	RMSE
Hybrid_GPD	0.965724	0.017194	0.971747	0.466526	0.007747
Normal_Gamma	0.99755	0.024987	0.977861	0.467671	0.0076
Normal_EVD	0.935869	0.024725	0.973789	0.547857	0.008253
EVD_lognorm	0.988356	0.024116	0.978971	0.422629	0.007247
EVD_Weibull	0.989687	0.030596	0.966731	0.770216	0.009717
Normal_Weibull	0.99658	0.025065	0.977833	0.469597	0.007615

Table 1: Tests of goodness-of-fit for each distribution on Tap 1

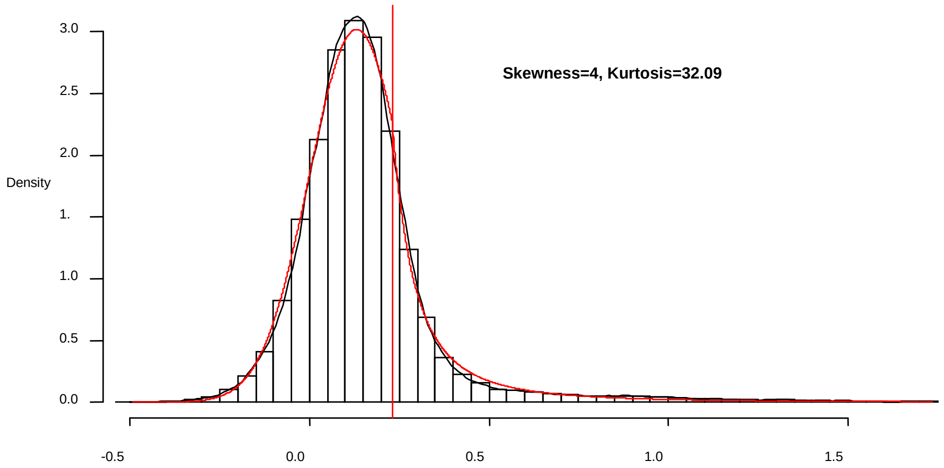

Fig 3: Histrogram and fitted curves for Tap 2

	Skewness= 4 and Kurtosis= 32.09				
	PPCC	K-S	R2	Chi2	RMSE
Hybrid_GPD	0.980687	0.007502	0.996213	0.207584	0.004117
Normal_Gamma	0.999642	0.014879	0.988241	0.775359	0.007606
Normal_EVD	0.817165	0.014017	0.987032	0.845275	0.008036
EVD_lognorm	0.982646	0.011328	0.996794	0.195879	0.003829
EVD_Weibull	0.997823	0.016941	0.990633	0.60967	0.006709
Normal_Weibull	0.999236	0.015041	0.987972	0.793639	0.007693

Table 2: Tests of goodness-of-fit for each distribution on Tap 2

In the analysis, the proportions of the tail parts are over 20% for both taps. The estimated thresholds are omitted but plotted in Firgure 2 and 3. From the Table 1, last three of the 5 measurements vote for EVD_lognorm while Hybrid_GPD has the smallest K-S values and PPCC of normal_gamma mixture is the largest one. Thus, overall EVD_lognorm may be a better choice for Tap1. In Table 2, we observe the similar results. EVD_lognorm again is the best based on the last three measurements. Based on the analysis, EVD_lognorm should be better choice for both taps. Without PPCC, Hybrid GPD has close performance to EVD_lognorm.

4 CONCLUSION

In this study, we majorly fit the pressure coefficients over the low-rise buildings with normal hybrid_GPD and two-component mixture distributions. We expected that hybrid GPD might fit the tail part of the data better than the mixtures, but the analysis results did not support that. It turns out EVD_lognorm mixture distribution has better fit among all 6 models. From Figure 2 and 3, we found that the estimated thresholds by hybrid GPD might be too small since the proportions of the tail parts for both taps are over 20%. Note that the extremes normally take small proportions of the data. This underestimated threshold might largely reduce the power of hybrid GPD method. On the other hand, we do observe that hybrid GPD has the smallest K-S values for both taps and hybrid GPD is only second to EVD_lognorm based on the last four measurements. Further study for hybrid GPD is guaranteed in the near future.

ACKNOWLEDGEMENT

The supports “the Fundamental Research Funds for the Central Universities (2682014CX060)” and “National 1000 Young Talents (China)” are greatly acknowledged.

REFERENCES

Akdağ, SA, Bagiorgas, HS, Mihalakakou, G (2010) “Use of two-component Weibull mixtures in the analysis of wind speed in the Eastern Mediterranean”. Applied Energy 87(8), 2566–2573

Benaglia T, Chauveau D, Hunter DR, Young D (2009c). “mixtools: An R Package for Analyzing Finite Mixture Models” Journal of Statistical Software, 32(6), 1-29.

Bordes L, Mottelet S, Vandekerkhove P (2006). “Semiparametric Estimation of a Two-Component Mixture Model”. The Annals of Statistics, 34(3), 1204-1232.

Carreau, J. and Y. Bengio (2009). “A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions”. IEEE transactions on neural networks 20 (7), 1087-1101.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer-Verlag: London.

Filliben, J.J. (1975). "The Probability Plot Correlation Coefficient Test for Normality", Technometrics, Vol. 17, No. 1, pp. 111-117

Ho, T.C.E., Surry, D., and Morrish, D.P. (2003). “NIST/TTU Cooperative Agreement–Windstorm Mitigation Initiative: Wind Tunnel Experiments on Generic Low Buildings.” Tech. Rep. BLWT-SS20-2003, The Boundary Layer Wind Tunnel Laboratory, The University of Western Ontario, London, Ontario, Canada.

Hu, Y. (2013). Extreme value mixture modelling: An R package and simulation study. MSc (Hons) thesis, University of Canterbury, New Zealand.

Jenkinson, A. F. (1955). “The frequency distribution of the annual maximum (or minimum) value of meteorological events. Quarterly” Journal of the Royal Meteorological Society 81 (348), 158-172.

Kollu et al. (2012) “Mixture probability distribution functions to model wind speed distributions” International Journal of Energy and Environmental Engineering, 3:27

MacDonald, A. E., C. J. Scarrott, and D. S. Lee (2013). “Boundary correction, consistency and robustness of kernel densities using extreme value theory”. Submitted.

MacDonald, A. E., C. J. Scarrott, D. S. Lee, B. Darlow, M. Reale, and G. Russell (2011). “A exible extreme value mixture model”. Computational Statistics and Data Analysis 55 (6), 2137-2157.

Reiss, R. and M. Thomas (2001). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkh□auser: Berlin.

Scarrott, C.J. and MacDonald, A. (2012). “A review of extreme value threshold estimation and uncertainty quantification”. Statistical Journal 10(1), 33-59.

von Mises, R. (1954). “La distribution de la plus grande de n valeurs”. American Mathematical Society: Providence RI II, 271-294