Tạp chí Khoa học và Công nghệ ***50*** *(1) (2012) 83-89*

XÁC ĐỊNH LUẬT PHÂN BỐ XÁC SUẤT CỦA DỮ LIỆU CHẤT LƯỢNG KHÔNG KHÍ ĐƯỢC QUAN TRẮC TẠI HÀ NỘI

Chử Thị Hồng Nhung[1], Nghiêm Trung Dũng[2]

[1]*Viện Khoa học và Kỹ thuật Môi trường, Trường Đại học Xây dựng*

[2]*Viện Khoa học và Công nghệ Môi trường, Trường Đại học Bách khoa Hà Nội*

[*]Email: *ngtdung-inest@mail.hut.edu.vn*

Đến Toà soạn ngày: 10/11/2010; Chấp nhận đăng ngày: 15/3/2012

1. ĐẶT VẤN ĐỀ

Một trong những nội dung quan trọng của công tác quản lí chất lượng không khí là hoạt động quan trắc. Và, một trong những thông tin quan trọng cần phải thu được từ các hoạt động quan trắc là tần suất mà nồng độ của một chất ô nhiễm không khí có thể vượt một giá trị cho trước, thường là ngưỡng cho phép của Quy chuẩn Việt Nam (QCVN), và xa hơn nữa là nồng độ cực đại có thể xảy ra đối với chất ô nhiễm đó. Trên thực tế, các chương trình quan trắc chất lượng không khí thường gián đoạn, nên số liệu thu được không liên tục. Ngay cả đối với các trạm quan trắc chất lượng không khí tự động cố định, vì nhiều lí do có thể như mất điện, sự cố kĩ thuật, ngừng để bảo dưỡng trạm … nên số liệu đo nhiều khi cũng không liên tục. Vì vậy, để có thể có được thông tin về tần suất vượt ngưỡng và giá trị nồng độ cực đại thì cần phải biết được luật phân bố xác suất của bộ số liệu quan trắc. Tuy nhiên, dữ liệu về nồng độ trung bình (từ 5 phút tới 24h) của các chất ô nhiễm không khí lại không tuân theo phân bố chuẩn (hay còn được gọi là phân bố Gauus) như rất nhiều đại lượng đo các thông số vật lí và hóa học khác mà tuân theo một số phân bố khác như lognormal hoặc Weibull [1, 2].

Trên thực tế, thông tin thu được từ các chương trình quan trắc chất lượng không khí hiện nay ở Việt Nam thường chỉ đơn giản là số lần đo hoặc số phần trăm của các điểm dữ liệu đo vượt ngưỡng của QCVN chứ không phải là tần suất vượt ngưỡng của QCVN [3]. Trong các tài liệu được công bố cũng chưa thấy có công trình nào về luật phân bố xác suất của dữ liệu chất lượng không khí ở Việt Nam nói chung và Hà Nội nói riêng. Nghiên cứu này, vì thế, đã được thực hiện nhằm xác định luật phân bố xác suất của dữ liệu chất lượng không khí được đo tại Hà Nội.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Phân bố xác suất thường được thể hiện qua hàm mật độ xác suất f(x) (Probability Density Function – PDF), biểu thị xác suất xuất hiện giá trị của đại lượng ngẫu nhiên X bằng với một giá trị x cụ thể nào đó theo luật phân bố xác suất [4]. Xác suất xuất hiện các giá trị của đại lượng ngẫu nhiên X nhỏ hơn hoặc bằng một giá trị cụ thể cho trước, khi đó, được biểu thị bằng hàm phân bố tần suất tích lũy F(x) (Cumulative Distribution Function – CDF) [5]:

$$F(x) = P\{X \le x\} = \int_{-\infty}^{x} f(x).dx$$

Nhưng trong thực tế thường dùng tần suất vượt P (thường chỉ được gọi tắt là tần suất) là xác suất xuất hiện các giá trị của đại lượng ngẫu nhiên X lớn hơn hoặc bằng một giá trị x cụ thể nào đó [5]:

$$P = P\{X \ge x\} = \int_{x}^{\infty} f(x).dx = 1 - P\{X \le x\} = 1 - F(x)$$

Phân bố Weibull: Là phân bố có hàm mật độ xác suất như sau [5]:

$$f_w(x) = (\frac{\gamma}{\alpha}).(\frac{x-\mu}{\alpha})^{\gamma-1}.\exp(-(\frac{x-\mu}{\alpha})^{\gamma}) \quad \text{với } x \ge \mu; \gamma, \alpha > 0$$

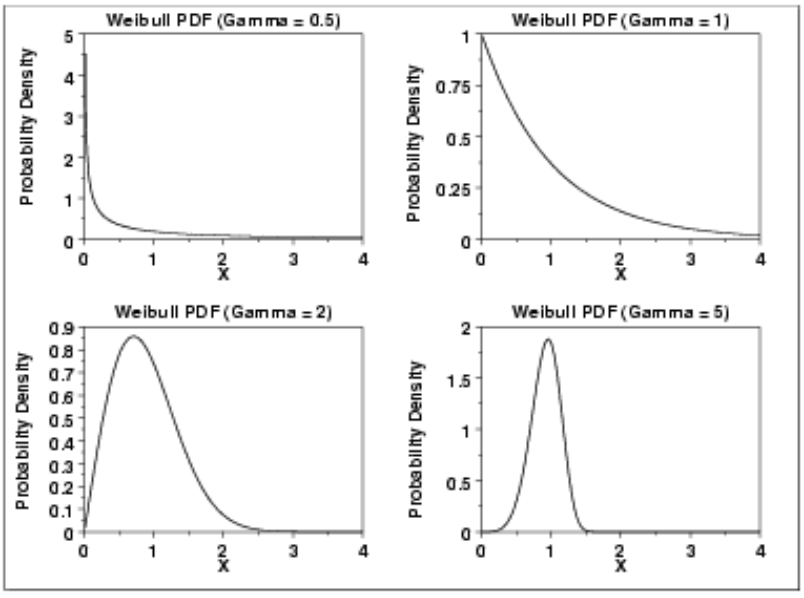
trong đó γ (gamma) là tham số hình dạng và α là tham số tỉ lệ của x. Nếu μ = 0 và α = 1 thì là phân bố Weibull chuẩn. Khi μ = 0 thì được gọi là phân bố Weibull hai tham số. Hàm mật độ phân bố phân bố Weibull chuẩn được biểu diễn như sau:

$$f(x) = \gamma.x^{(\gamma-1)}.\exp(-(x^{\gamma}) \text{ với } x \ge 0; \gamma > 0$$

Công thức hàm phân bố xác suất tích lũy của phân bố Weibull là:

$$F(x) = 1 - \exp(-(x^{\gamma})) \text{ với } x \ge 0; \gamma > 0$$

Một số dạng đường cong của hàm mật độ xác suất của phân bố Weibull được trình bày trên hình 1.



Hình 1. Biểu diễn hàm mật độ xác suất của phân bố Weibull

Phân bố lognormal: Là phân bố có hàm mật độ xác suất như sau [6]:

$$f_L(x) = \frac{1}{(x-\theta).\sigma.\sqrt{2\pi}}.\exp(-\frac{(\ln(x-\theta)-\ln m)^2}{2\sigma^2}) \text{ với } x \geq \theta; m, \sigma > 0$$

trong đó, σ là tham số hình dạng, θ là tham số vị trí và m là tham số tỉ lệ. Trong trường hợp θ = 0 và m = 1 thì có phân bố lognormal chuẩn. Trường hợp θ = 0 thì có phân bố lognormal hai tham số. Phân bố lognormal chuẩn có công thức sau:

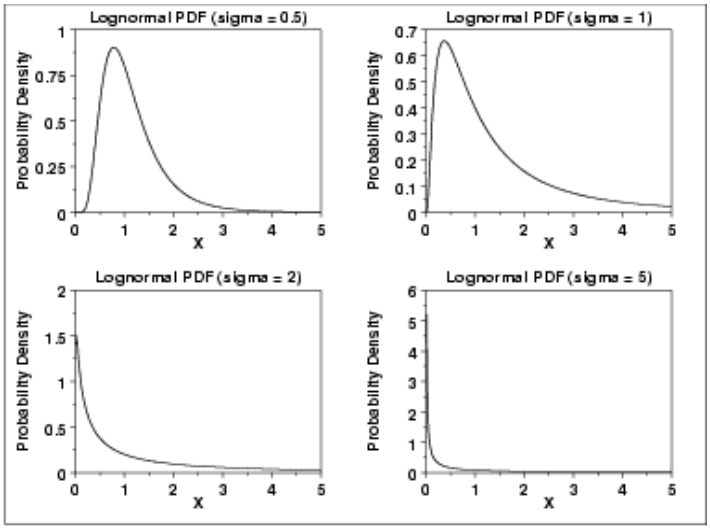
$$f(x) = \frac{1}{x.\sigma.\sqrt{2\pi}}.\exp(-(\ln x)^2 / 2\sigma^2) \text{ với } x \geq 0; \sigma > 0 .$$

Hàm phân bố tần suất tích lũy của phân bố lognormal là:

$$F(x) = \Phi(\frac{\ln(x)}{\sigma}) \text{ với } x \geq 0; \sigma > 0$$

trong đó, Φ là hàm phân bố xác suất tích lũy của phân bố chuẩn.

Một số dạng đường cong của hàm mật độ xác suất của phân bố lognormal được trình bày trên hình 2.



Hình 2. Biểu diễn hàm mật độ xác suất của phân bố Log-normal

Việc xác định luật phân bố xác suất của dữ liệu quan trắc chất lượng không khí được thực hiện thông qua nghiên cứu nhận dạng và xác định được các tham số của phân bố xác suất dựa trên dữ liệu chất lượng không khí của Hà Nội với sự hỗ trợ của phầm mềm thống kê SPSS.

Dữ liệu được sử dụng trong nghiên cứu này là nồng độ trung bình giờ của các thông số ô nhiễm gồm SO_2, NO_2, CO, O_3, TSP và PM_{10} thu được tại trạm quan trắc không khí tự động Láng, Hà Nội trong khoảng thời gian gần 8 năm, từ 7/2002 đến 5/2010. Như vậy mỗi thông số có khoảng 8760 điểm dữ liệu trong một năm và tổng khoảng 832.200 điểm dữ liệu nếu tính cho toàn bộ khoảng thời gian trên (khoảng 95 tháng). Trước khi sử dụng để để nhận dạng phân bố, các điểm ngoại biên đã được loại bỏ dựa trên giá trị phân vị.

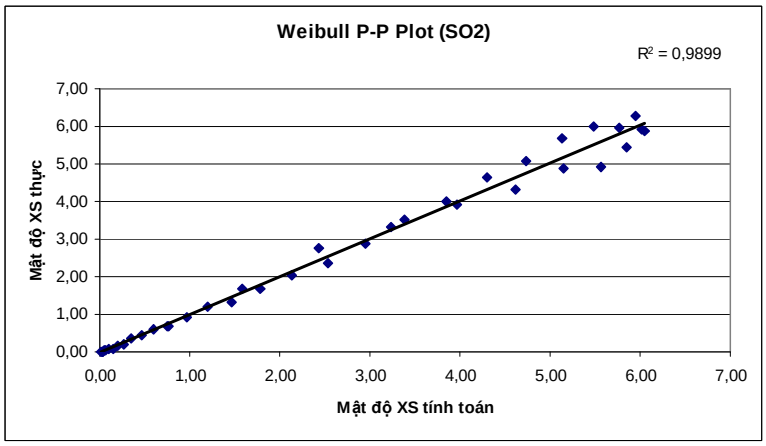
Phần mềm SPSS có công cụ để kiểm tra dạng phân bố của dữ liệu, đó là sử dụng P-P Plot. SPSS sẽ đưa ra một mô hình sao cho giá trị dự báo của mô hình có sai số là nhỏ nhất so với giá

trị của bộ dữ liệu thực tế có được. Khi hàm mật độ xác suất được xác định thì hàm phân bố tần suất tích lũy hay tần suất vượt ngưỡng cũng được xác định.

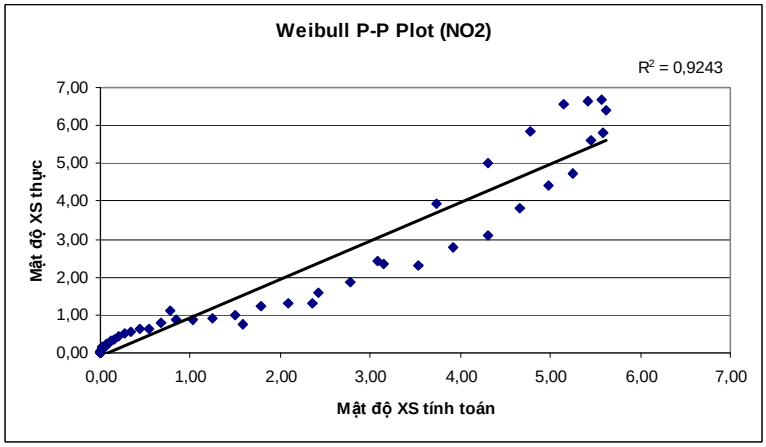
3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Nhận dạng luật phân bố xác suất

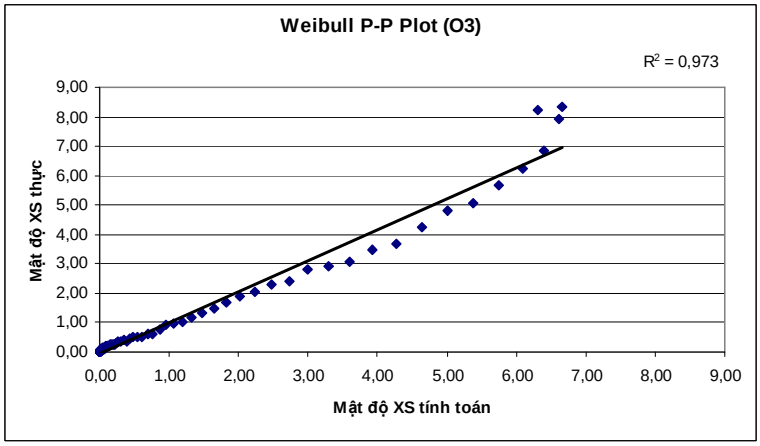
Với sự hỗ trợ của công cụ SPSS, đã kiểm tra, nhận dạng sự phù hợp của dữ liệu theo các dạng phân bố xác suất khác nhau. Kết quả cho thấy rằng, nồng độ trung bình giờ của các chất ô nhiễm nói trên tuân theo luật phân bố Weibull. Điều này có thể được khẳng định do mức độ tương quan cao giữa giá trị nồng độ tính toán theo mô hình và giá trị nồng độ đo được của các chất ô nhiễm với hệ số tương quan mẫu R^2 đều lớn hơn 0,9. Biểu đồ P-P Plot của các chất ô nhiễm SO_2, NO_2, O_3, được trình bày trên các hình 3 - 5, cho thấy rõ điều đó.



Hình 3. Biểu đồ Weibull P-P Plot của SO_2

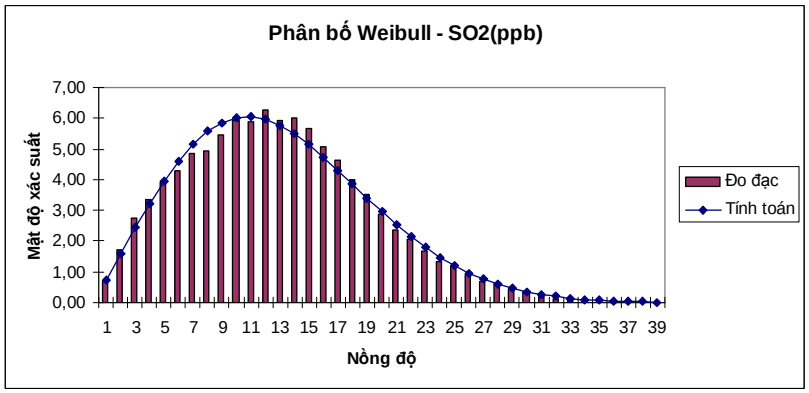


Hình 4. Biểu đồ Weibull P-P Plot của NO_2

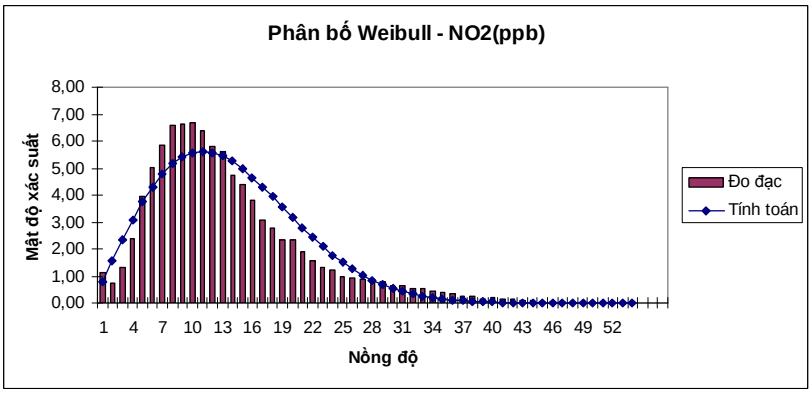


Hình 5. Biểu đồ Weibull P-P Plot của O_3

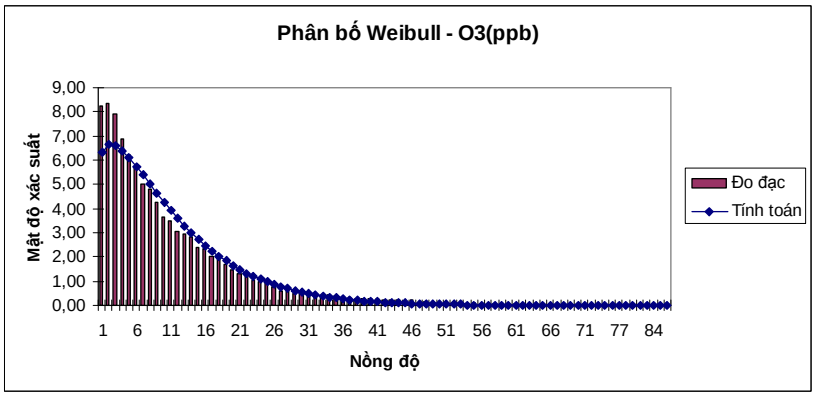
Sự tương quan tốt cũng có thể được thấy rõ trên các Hình 6-8, trình bày dạng đường cong thu được theo tính toán (Weibull) và theo số liệu thực nghiệm (đo đạc) của hàm mật độ xác suất. Trong đó, mật độ xác suất có đơn vị là %, còn đơn vị của nồng độ các chất ô nhiễm là ppb.



Hình 6. Phân bố Weibull của SO_2



Hình 7. Phân bố Weibull của NO_2



Hình 8. Phân bố Weibull của O_3

3.2. Xác định các tham số của phân bố Weibull

Các tham số tỉ lệ và tham số hình dạng của phân bố Weibull đối với bộ dữ liệu sử dụng đã được xác định và được trình bày trong bảng 1.

Bảng 1. Tham số của hàm phân bố Weibull đối với các chất ô nhiễm

Tham số	**SO_2_ppb**	**NO_2_ppb**	**CO_ppm**	**O_3_ppb**	**TSP_ µg/m^3**	**PM_{10}_µg/m^3**
Tham số tỉ lệ α	14,644	15,402	0,800	11,434	102,357	79,473
Tham số hình dạng γ	2,097	2,032	1,826	1,177	1,992	1,861

Tuy nhiên, biết được hình dạng phân bố dữ liệu chỉ là bước đầu để hiểu rõ hơn về đặc điểm của bộ dữ liệu. Mục tiêu quan trọng hơn là xác định hàm mật độ xác suất và hàm phân bố tần suất tích lũy để từ đó xác định được tần suất vượt ngưỡng và nồng độ cực đại có thể xảy ra đối với một chất ô nhiễm. Với việc xác định được các tham số như được trình bày trong Bảng 1 thì hàm mật độ xác suất và hàm phân bố tần suất tích lũy của dữ liệu về nồng độ trung bình giờ của các chất ô nhiễm nói trên đã được xác định

3.3. Khả năng áp dụng phân bố Weibull cho Hà Nội:

Kết quả thu được về sự tuân theo phân bố Weibull là dựa trên dữ liệu quan trắc chất lượng không khí tại trạm Láng, Hà Nội. Tuy nhiên theo [7], chỉ có một vài dạng nguồn thải chính chi phối chất lượng không khí ở địa bàn thành phố Hà Nội (cũ). Vì vậy phân bố Weibull cũng có thể đúng cho các khu vực khác của thành phố Hà Nội. Hay nói cách khác, có thể áp dụng phân bố Weibull cho việc xử lí dữ liệu quan trắc chất lượng không khí ở Hà Nội.

4. KẾT LUẬN

Kết quả nghiên cứu thu được đã cho thấy rằng, nồng độ trung bình giờ của các chất ô nhiễm không khí quan trắc tại Hà Nội tuân theo phân bố Weibull. Từ đó, đã xác định được hàm mật độ xác suất và hàm phân phối xác suất tích lũy của bộ dữ liệu này. Điều này rất có ý nghĩa đối với công tác quản lí chất lượng không khí ở Hà Nội vì nó cho phép tính được tần suất mà nồng độ trung bình của một chất ô nhiễm không khí có thể vượt một giá trị cụ thể nào đó, ví dụ ngưỡng cho phép của QCVN. Biết được luật phân bố xác suất cũng cho phép xác định được nồng độ cực đại có thể xảy ra đối với một chất ô nhiễm nào đó cho dù chúng ta không có kết quả đo của giá trị cực đại đó. Phương pháp luận sử dụng trong nghiên cứu này có thể được áp dụng cho các địa phương khác của nước ta để xác định luật phân bố xác suất của dữ liệu chất lượng không khí. Và hy vọng điều này có thể góp phần vào việc nâng cao chất lượng của công tác xử lí dữ liệu quan trắc chất lượng không khí hiện nay ở Việt Nam.

Lời cảm ơn. Các tác giả xin chân thành cảm ơn Trung tâm Mạng lưới Khí tượng, Thủy văn và Môi trường đã cung cấp số liệu quan trắc chất lượng không khí cho nghiên cứu này.

TÀI LIỆU THAM KHẢO

1. John H. Seinfeld and Spyros N. Pandis. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. Second Edition, John Wiley & Sons, Inc., 2006.
2. Kenneth E. Noll and Terry L. Miller. Air Monitoring Survey Design. Ann Arbor Science Publishers Inc., 1977.
3. Bộ Tài nguyên và Môi trường. Báo cáo môi trường quốc gia năm 2007: Môi trường không khí đô thị Việt Nam, 2007.
4. Tống Đình Quỳ. Giáo trình xác suất thống kê. Nhà xuất bản Đại học Quốc gia Hà Nội, 2003.
5. Nghiêm Tiến Lam (2008). Tính toán tần suất theo phân bố Weibull. Khoa Kỹ thuật Biển, Đại học Thủy lợi, http://coastal.wru.edu.vn/Thu_vien /cepg/Phan%20bo%20tan%20suat%20Weibull.pdf (truy cập ngày 15/10/2010)
6. SEMATECH, NIST (2010), Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm.
7. Nghiêm Trung Dũng. Nghiên cứu mức độ phát thải và lan truyền của các hydrocarbon thơm đa vòng (PAHs) tại Hà Nội, Luận án tiến sỹ kĩ thuật, Trường Đại học Bách khoa Hà Nội, 2005.

SUMMARY

DETERMINATION OF THE PROBABILITY DISTRIBUTION OF AIR QUALITY DATA MONITORED AT HANOI

A study to determine the probability distribution of air quality data measured at Hanoi was conducted using the statistical software of SPSS. Data used are hourly average concentrations of air pollutants including SO_2, NO_2, CO, O_3, TSP and PM_{10} collected from Lang automatic ambient air quality monitoring station, Hanoi for a period of about 8 years, from July 2002 to May 2010. Obtained results show that the Weibull distribution was fitted to the data of hourly average concentrations of these air pollutants. As a result, Probability Density Function (PDF), f(x) and Cumulative Distribution Function (CDF), F(x) for these set of data were determined.