Air Pollution Statistics Based on Superstatistics

Authors: Wang Huilin, Deng Weibing

Date: 2024-09-26T00:00:00+00:00

Abstract

Air pollution caused by carbon monoxide (CO) and sulfur dioxide (SO2) represents a major challenge in large metropolitan areas, as these two substances pose serious health risks. However, the statistical characteristics of air pollutants are not yet fully elucidated. This paper employs methodologies from non-equilibrium statistical mechanics to construct a superstatistical model tailored for the statistical analysis of air pollution. Concurrently, we analyze time series of carbon monoxide (CO) and sulfur dioxide (SO2) concentrations recorded in Wuhan. We find that the probability distributions of concentrations follow q-Gaussian distributions, and the dynamics can be well described by 2 superstatistics. Our findings enable precise risk assessment of high-pollution events and provide a foundation for developing mitigation strategies.

Full Text

Preamble

Air Pollution Statistics Based on Superstatistics

Wang Huilin[1,2]), Deng Weibing[1,2])†

[1]) (Institute of Particle Physics, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China)
[2]) (Key Laboratory of Quark and Lepton Physics (MOE), Central China Normal University, Wuhan 430079, China)

Air pollution caused by carbon monoxide (CO) and sulfur dioxide (SO_2) represents a major challenge for large metropolitan areas due to the severe adverse health effects associated with these pollutants. However, the statistical properties of air pollutants remain incompletely understood. In this paper, we adapt methods from non-equilibrium statistical mechanics to construct a superstatistical model appropriate for air pollution statistics. Simultaneously, we analyze time series data of CO and SO_2 concentrations recorded in Wuhan. We find

that the probability distributions of concentrations follow q-Gaussian distributions, and the dynamics can be well described by ^{2} superstatistics. Our results enable precise risk estimation for high-pollution scenarios and pave the way for developing effective mitigation strategies.

Keywords: superstatistics, time series, air pollution, non-equilibrium statistical mechanics
PACS: 05.20.Dd, 95.60.Bh, 92.60.Sz, 02.50.-r

These authors contributed equally.

† Corresponding author. E-mail: wdeng@mail.ccnu.edu.cn (Deng Weibing)
First author. E-mail: 15623826232@163.com (Wang Huilin)
Project supported by the National Natural Science Foundation of China (Grant Nos. 11505071, 61702207, 61873104)

---

1 Introduction

Superstatistics provides a general framework for describing the dynamics of complex non-equilibrium systems with well-separated time scales. These models generate heavy-tailed non-Gaussian distributions through a simple mechanism: the superposition of simple distributions whose relevant parameters fluctuate as random variables over longer time scales. Originally developed for turbulence modeling, superstatistics has since been applied to diverse physical systems including plasma physics, Ising models, cosmic ray physics, self-gravitating systems, solar wind, high-energy scattering processes, ultra-cold gases, and non-Gaussian diffusion in small complex systems. The framework has also proven successful in entirely different domains such as power grid frequency modeling, wind statistics, air pollution, bacterial DNA, financial time series, rainfall statistics, and train delays. Comprehensive review articles provide contemporary introductions to superstatistics and non-Gaussian diffusion.

In all these cases, a fundamental simple distribution—typically Gaussian or exponential—has been identified to explain the heavy tails observed in aggregated marginal distributions when fluctuating parameters are accounted for. These tails typically decay as power laws. Here we apply superstatistical analysis to a critically important topic: the spatiotemporal dynamics of air pollution in major cities. While previous studies have examined different gases, we focus primarily on CO and SO_2 as primary examples, though the methodology applies equally to other pollutants.

Sulfur dioxide (SO_2) and carbon monoxide (CO) are common yet hazardous air pollutants that pose serious threats to both human health and the environment. SO_2 irritates the respiratory tract, causing coughing, sore throat, bronchitis, and asthma, while potentially increasing the risk of heart attacks and strokes.

Furthermore, SO_2 transforms into sulfuric acid in the atmosphere, forming acid rain that corrodes buildings and damages crops and forest ecosystems. CO, conversely, binds to hemoglobin, reducing the blood's oxygen-carrying capacity and causing hypoxia symptoms including headaches, dizziness, and nausea; severe cases can lead to coma or death. Although CO's direct environmental impact is relatively minor, it oxidizes to carbon dioxide in the atmosphere, contributing to the greenhouse effect. Reducing emissions of both SO_2 and CO is therefore essential for protecting public health and the environment.

The data for this analysis were obtained from the publicly accessible Zhenqiwang platform, which provides urban pollutant level readings for Wuhan. We analyzed hourly interval data. Examination of the measured concentration time series for both CO and SO_2 reveals variations across multiple time scales (see [Figure 1: see original paper]).

2.1 Theoretical Model

The fundamental concept of superstatistics is that a long time series exhibiting complex, heavy-tailed probability distributions actually represents an ensemble of many shorter time series, each producing a simple non-heavy-tailed distribution. This approach has been successfully applied to numerous types of complex systems. As a first step in superstatistical time series analysis, we must extract a long time scale T over which we locally observe simple distributions.

Assuming the time series is approximately Gaussian within short temporal slices, the kurtosis of local snapshots should be $\kappa_{\mathrm{Gaussian}} = 3$. In contrast, the fully aggregated time series exhibits higher kurtosis κ. To determine T, we test different time window sizes Δt and compute the local average kurtosis as follows:

$$\kappa(\Delta t) = \frac{\langle (u-\bar{u})^4 \rangle_{t_0,\Delta t}}{\langle (u-\bar{u})^2 \rangle^2_{t_0,\Delta t}}$$

where $t_{\max}$ is the length of the time series and $\langle \cdots \rangle_{t_0,\Delta t}$ denotes the expectation over a time slice of length Δt starting at t_0. The long time scale is then determined by the condition $\kappa(T) = \kappa_{\mathrm{Gaussian}}$, meaning the average kurtosis over windows of length T equals the Gaussian value of 3. After determining T, we can segment the time series into samples of length T, obtaining a collection of approximately locally Gaussian distributions, each characterized by a different inverse variance parameter β.

If these β values follow a χ^2 distribution, it can be expressed as:

$$f(\beta) = \frac{1}{\Gamma(n/2)} \left(\frac{n}{2\beta_0} \right)^{n/2} \beta^{n/2-1} e^{-n\beta/(2\beta_0)}$$

where n represents the degrees of freedom and β_0 is the mean. Analytical integration then yields a q-Gaussian distribution for the aggregated statistics. Alternatively, the β distribution may be described by other distributions such as inverse-χ^2 or log-normal. In these cases, the resulting marginal distributions differ (though they are often well-approximated by q-Gaussians). Whether a given time series follows χ^2, inverse-χ^2, or log-normal superstatistics depends on the actual distribution of β.

Since superstatistics was originally derived for temperature fluctuations, β is often interpreted as an inverse temperature related to local kinetic energy in the system. More generally, however, it simply represents a fluctuating inverse variance parameter for a given time series. For general superstatistics, we expect to observe q-Gaussian probability density functions of the form:

$$p(q, b, \mu) = C_q \left[1 + (1-q)(-b(x-\mu)^2)\right]^{\frac{1}{1-q}}$$

where C_q is a normalization constant, μ is a shift parameter, q is the shape parameter (also called the entropic index), and b is a scale parameter proportional to the expectation $\langle\beta\rangle$ derived from the distribution in the previous equation. In the limit $q \to 1$, the q-Gaussian distribution reduces to a Gaussian with variance $1/2b$. A specialized monograph on q-statistics applications in hydraulic engineering provides further details. Note that the superstatistical distributions described here may arise from Gaussian processes with time-dependent standard deviations—that is, superpositions of simple Gaussian distributions over long times.

The long time scale T characterizes the time scale over which the underlying stochastic process varies, while the short time scale τ gives the system's relaxation time to local equilibrium, defined through the decaying autocorrelation of the original time series approximated as $c \sim e^{-t/\tau}$. For the superstatistical approach to remain valid, we must assume $\tau \ll T$ to ensure the system can always relax to a new equilibrium.

2.2 Superstatistical Analysis of Air Pollution

By examining fluctuations in two time series of Wuhan air pollution data (CO and SO_2), we first note the highly non-Gaussian nature of the distributions, which can be captured by q-Gaussian distributions (see [Figure 2: see original paper]). The q-Gaussian clearly provides a better fit to the empirical data than the Gaussian distribution, as shown in the figure and equation (3). To investigate the origin of these non-Gaussian distributions, we proceed with superstatistical analysis: assuming the non-Gaussian distributions arise from locally Gaussian distributions, we extract a long time scale T over which the local distributions are Gaussian. We identify this long scale as the time window where the data's average kurtosis satisfies $\kappa(T) = \kappa_{\text{Gaussian}} = 3$ (see [Figure 3: see original paper]). For CO concentration analysis, we observe a long time scale of $T_{\mathrm{CO}} = 40$ hours, while for SO_2, we find $T_{\mathrm{SO2}} = 17$ hours.

Our method for testing fluctuations over length scales t where local distributions approximate Gaussians involves segmenting the hourly data into windows: 40 hours for CO and 17 hours for SO_2 (see [Figure 3: see original paper]). [Figure 4: see original paper] presents the local distributions for both pollutants, which can be approximated by Gaussian fits with probability density functions $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, where μ is the location parameter and σ^2 is the variance.

[Figure 4: see original paper] illustrates the procedure for calculating T. Once T is determined, we obtain the parameter distribution $f(\beta)$ from histograms by setting $\beta = 1/\sigma^2$, where σ^2 is the variance of the data. As shown in [Figure 5: see original paper], both CO and SO_2 exhibit β distributions that follow χ^2 distributions. For CO, maximum likelihood estimation yields a χ^2 distribution with $k = 2.5$, while for SO_2, a χ^2 distribution with $k = 0.6992$ provides an excellent fit.

---

3 Discussion

Urban pollution policies often focus on threshold values and compliance monitoring, which frequently results in pollution levels clustering just below critical values. However, reducing overall pollutant exposure may be more rational in many contexts. We therefore outline how knowledge of the probability density function enables estimation of threshold crossings and total exposure.

Setting concentration thresholds represents an effective policy tool, yet data coverage is often incomplete, particularly when only a few months of data are available for the period of interest, making it difficult to estimate violation frequencies. Once we determine the distribution $P(u)$ of concentration levels u, we can easily compute the number of days exceeding a threshold as:

$$N_{u>u_{\text{threshold}}} = 365 \int_{u_{\text{threshold}}}^{\infty} p(u)\, du$$

This integral can be readily evaluated numerically. Using the q-Gaussian distribution reveals that the number of threshold violations depends significantly on parameters q and μ, which must be determined from the dataset but can be compared across locations and time periods. These parameters separately characterize two distinct aspects: the incidence of (rare) extreme events, captured by q, and the overall variability of pollutant concentrations, captured by μ.

As a complement to thresholds, we can focus on total pollution exposure (TPE)—the average amount of pollutant to which citizens are exposed—which can also be easily calculated from the distribution:

$$\text{TPE} = 365 \int_0^{\infty} u \cdot p(u)\, du$$

In summary, we have demonstrated that the superstatistical approach, originally introduced in non-equilibrium statistical mechanics and applied to turbulence and other complex systems, can quantitatively model air pollution statistics. We find that simple superstatistical models agree remarkably well with empirical pollution data. A key result is that different pollutants obey the same type of superstatistics (CO and SO_2 in our case). Once the precise superstatistical model for a specific pollutant is established, precise risk estimates for high-pollution scenarios can be obtained by integrating the tails of the probability distribution above a given threshold. These estimates facilitate tailored threshold design for effective urban pollution policies. The superstatistical framework presented here can be readily generalized and extended—for instance, by investigating precise superstatistical models for additional pollutants and exploring alternative local distributions, potentially linking these statistics to the physical and chemical properties of pollutants.

References

[1] Beck C, Cohen E G 2003 *Physica A: Statistical Mechanics and its Applications* 322 267
[2] Beck C 2007 *Physical Review Letters* 98 064502
[3] Livadiotis G 2017 *Kappa Distributions: Theory and Applications in Plasmas* (Elsevier)
[4] Cheraghalizadeh J, Seifi M, Ebadi Z, Mohammadzadeh H, Najafi M 2021 *Physical Review E* 103
[5] Yalcin G C, Beck C 2018 *Scientific Reports* 8 1764
[6] Ourabah K 2020 *Physical Review D* 102 043017
[7] Livadiotis G, Desai M, Wilson L 2018 *The Astrophysical Journal* 853 142
[8] Ayala A, Hernández-Ortiz S, Hernandez L A, Knapp-Pérez V, Zamora R 2020 *Physical Review D*
[9] Rouse I, Willitsch S 2017 *Physical Review Letters* 118 143401
[10] Itto Y, Beck C 2021 *Journal of the Royal Society Interface* 18 20200927
[11] Schäfer B, Beck C, Aihara K, Witthaut D, Timme M 2018 *Nature Energy* 3 119
[12] Weber J, Reyers M, Beck C, Timme M, Pinto J G, Witthaut D, Schäfer B 2019 *Scientific Reports* 9
[13] Williams G, Schäfer B, Beck C 2020 *Physical Review Research* 2 013019
[14] Bogachev M I, Markelov O A, Kayumov A R, Bunde A 2017 *Scientific Reports* 7 43034
[15] Gidea M, Katz Y 2018 *Physica A: Statistical Mechanics and its Applications* 491 820

[16] Uchiyama Y, Kadoya T 2019 *Physica A: Statistical Mechanics and its Applications* 526 120930
[17] De Michele C, Avanzi F 2018 *Scientific Reports* 8 14204
[18] Briggs K, Beck C 2007 *Physica A: Statistical Mechanics and its Applications* 378 498
[19] Metzler R 2020 *The European Physical Journal Special Topics* 229 711
[20] Beck C, Cohen E G, Swinney H L 2005 *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 72 056133
[21] Beck C 2001 *Physical Review Letters* 87 180601
[22] Singh V P 2016 *Introduction to Tsallis Entropy Theory in Water Engineering* (CRC Press)