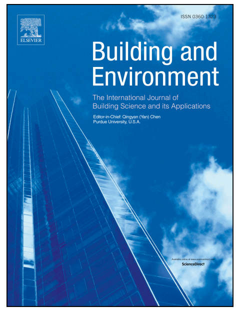
Journal Pre-proof



Sampling method for long-term monitoring of indoor environmental quality in residential buildings

Huimin Yao, Xiaojie Cheng, Shen Wei, Yuling Lv, Ang Li, Xiong Shen

PII: S0360-1323(22)00207-4

DOI: https://doi.org/10.1016/j.buildenv.2022.108965

Reference: BAE 108965

To appear in: *Building and Environment*

Received Date: 8 December 2021

Revised Date: 2 March 2022

Accepted Date: 4 March 2022

Please cite this article as: Yao H, Cheng X, Wei S, Lv Y, Li A, Shen X, Sampling method for long-term monitoring of indoor environmental quality in residential buildings, *Building and Environment* (2022), doi: https://doi.org/10.1016/j.buildenv.2022.108965.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Ltd.

Journal Pre-proof

Sampling method for long-term monitoring of indoor environmental quality in residential buildings

Huimin Yao[1], Xiaojie Cheng[1], Shen Wei[2], Yuling Lv[1], Ang Li[1], Xiong Shen*[1]

[1] Tianjin Key Lab of Indoor Air Environmental Quality Control, School of Environmental Science and Engineering, Tianjin University, Tianjin 300072, China

[2] The Bartlett School of Sustainable Construction, University College London (UCL), 1-19 Torrington Place, London WC1E 7HB, United Kingdom

E-mail for the corresponding author: shenxiong@tju.edu.cn

Abstract:

The data collected during long-term monitoring (LTM) of indoor environmental quality (IEQ) can reflect occupants' exposure to contaminants and can be used to improve thermal comfort. As there are large differences among existing guidelines for IEQ monitoring of dwellings, it is important to identify a sampling method that balances data accuracy, sample size and cost. This paper reports the major findings that developed a systematic approach to determining the sample method for IEQ monitoring. In the study, LTM was carried out in 13 naturally ventilated urban residences in Kunming, China. We proposed the continuous sampling strategy (CSS) and discrete sampling strategy (DSS). Descriptive statistics was used to evaluate the performances of both strategies, and it was found that DSS could obtain more accurate data than CSS. Next, an algorithm was developed for calculating the optimal sampling frequencies for different parameters based on the Pearson correlation coefficient. We evaluated the required number of dwellings(RND) for various parameters that satisfied the statistical

Journal Pre-proof

confidence in Kunming and other four cities of China. We found that with the increase in the household number in one city, the RND will reach to a critical threshold and no longer increase anymore. Using this threshold and the simple random sampling principle, we also provide guidance for determining the RND for IEQ monitoring in residence. It is expected that the results of this study will facilitate the selection of sampling method for similar studies in the future, with reduced manpower and consumption but a representative sample.

Keywords: Indoor environmental quality, Sampling strategy, Long-term monitoring, Pearson correlation coefficient, Sampling frequency

Nomenclature

TVOC	Total volatile organic compound
CO_2	Carbon dioxide
LTM	Long-term Monitoring
IEQ	Indoor Environmental Quality
CSS	Continuous Sampling Strategy
DSS	Discrete Sampling Strategy
PCC	Pearson Correlation Coefficient
RND	Required number of dwellings.

1. Introduction

According to the National Human Activity Pattern Survey in the United States [1],

Journal Pre-proof

Journal Pre-proof

people spend about 87% of their time indoors, and this percentage is increasing. Indoor air quality (IAQ) and thermal comfort are key factors in the health of building occupants [2], because poor indoor environmental quality (IEQ) may lead to respiratory diseases or sick building syndrome [3]. In addition, ensuring clean and comfortable indoor air is an important public health goal [4]. The evaluation and analysis of IEQ commonly rely on field monitoring in actual buildings. Therefore, accurate monitoring becomes necessary for solving indoor pollution problems and improving people's overall well-being [5].

The analysis of IEQ normally involves both short-term monitoring (STM) and long-term monitoring (LTM). Because IEQ parameters can be affected by outdoor weather (daily and seasonally), human behavior and building attributes [6-8], incomplete and varying IEQ measurement results are common [9]. Because of the characteristics of IEQ parameters, it is necessary to monitor the residential environment for a longer time. In comparison with STM data, the data collected by LTM have several advantages: 1) LTM data enable the detection of peak concentration values [10]; 2) the data reflect real-time exposure to indoor pollutants [11]; and 3) the data can be used as feedback signals for real-time pollutant control [7]. Therefore, LTM techniques have been widely adopted to capture critical indoor environmental parameters, primarily temperature, humidity, as well as concentration of carbon dioxide(CO_2), PM2.5, formaldehyde and total volatile organic compound (TVOC) [12]. In the collection of data by long-term monitoring, the quality of data will determine whether the data correctly reveal the basic conditions of relevant indoor parameters and represents the

Journal Pre-proof

characteristics of human exposure to pollutants [13].

To reduce the effort and cost entailed by IEQ-related studies, we need to consider decisions about both data sampling strategies and sampling frequency [14]. Hui et al. [15] used CO_2 as a reference to evaluate existing and proposed sampling schemes for indoor pollutant concentration in terms of the necessary sampling time and sampling point density and the probable errors induced at certain confidence levels of the measurement. In Hong Kong, continuous sampling [16] for a measurement period of 8 h has been generally adopted to determine the average pollutant concentration in a workplace. Since building-related contaminants normally peak in the morning in workplaces, and occupant-related contaminants in the afternoon, Mui et al. [17] proposed a new sampling strategy that uses the average concentration of two random measurement samples. According to the results, as compared to the typical 8-h measurement period of a continuous sampling method in Hong Kong, the measurement time with the new method could be reduced by up to 30%. Christopher [18] analyzed a rich data set and found that indoor particle events tend to be brief, intermittent, and highly variable. Hence, to characterize sources of PM in indoor environments, he used both continuous and time-integrated sampling instruments to simultaneously measure indoor/outdoor (I/O) particle mass concentration. Based on the different characteristics of indoor air parameters, previous studies have often used different monitoring and sampling methods, but a unified conclusion has not been formed. Furthermore, the sampling strategy and sampling frequency usually depend only on the sampling precision of the instrument and the labor cost [19]. A comprehensive IEQ data sampling

Journal Pre-proof

Journal Pre-proof

method along with IAQ audit methodology for buildings should be established for identification of indoor air problems.

Both sampling frequency and sampling strategies have been addressed in existing standards from different countries and regions, as summarized in Table 1. These standards recommend sampling frequencies for various major IEQ parameters. For example, for formaldehyde, China's GB/T18883 standard [20] recommends taking the daily average. The list in Table 1 also includes different sampling periods, either real time (WHO [21]) or an 8-hour average (US-EPA [22]). The sampling periods for formaldehyde, PM2.5 and VOC also differ in these standards. According to the Indoor Air Sampling and Evaluation Guide [22] formulated by the U.S. Environmental Protection Agency, monitoring strategies for indoor air pollutants should be evaluated in terms of the pollutants' exposure level, exposure time, pollutant toxicity and pollutant concentration. The above standards have usually relied on previous experimental results and experience in establishing the threshold value [23]. But because of the different background of the experiment, different countries and regions have very different requirements for the same IEQ parameter. Furthermore, according to WHO, there is insufficient evidence that indoor pollutants do not cause adverse effects when they fall below the thresholds in the standards [21]. Therefore, short-term measurement and the threshold concentrations proposed by the standards may not be enough. In order to study IEQ and ensure the health and thermal comfort of occupants, we may also need to study the sampling frequency and sampling strategies of LTM.

Journal Pre-proof

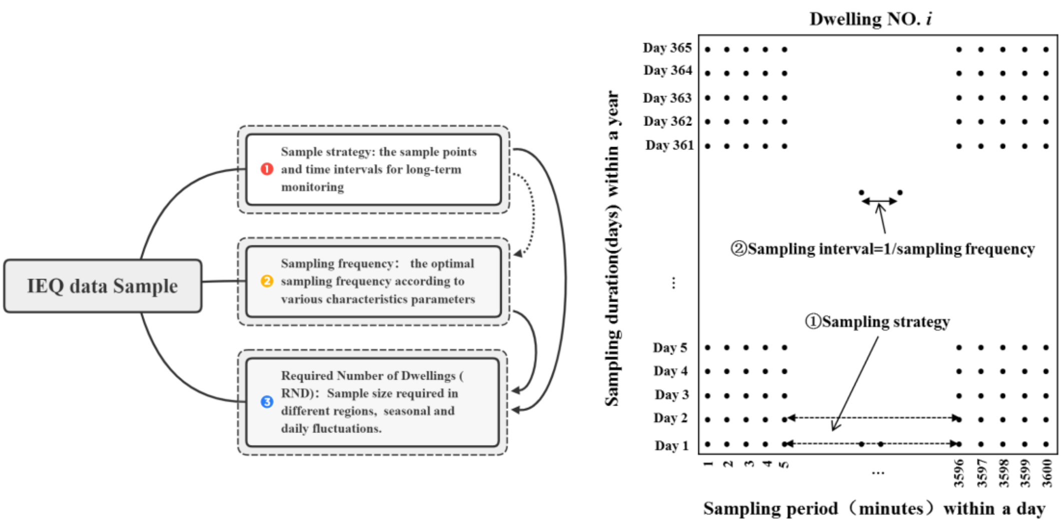
Journal Pre-proof

Table 1 Varying regulations for threshold concentration and sampling period of IEQ parameters in residential buildings

Index		Sampling period	Threshold concentration
Carbon dioxide (ppm)			
China-GB/T18883[20]		24-hour average	1000
China HK [24]	Good level	8-hour average	1000
	Excellent level	8-hour average	800
WHO-Europe[25]		1-hour average	900
Singapore[26]		8-hour average	1000
NIOSH[27]		8-hour average	5000
		15-min average	30000
Canada[28]		real time	3500
UK[29]		15-min average	15000
		5-min average	5000
Australia[30]		15-min average	30000
US-EPA[22]		real time	800
PM2.5(μg/m³)			
WHO[21]		annual average	10
		24-hour average	25
Canada-EGR		1-hour average	100
		real time	40
China- JGJ/T 309		24-hour average	75
US-EPA		24-hour average	65
Formaldehyde (μg/m³)			
Canada-EGR		real time	120
WHO		30-min average	100
		real time	200
Singapore		8-hour average	120
NIOSH		15-min average	0.1ppm
UK		15-min average	2500
Australia		15-min average	2500
US-EPA		8-hour average	920
China-GB/T18883		24-hour average	100
China-GB/T50325[31]		24-hour average	100
China HK-IAQC	Good level	30-min average	100
	Excellent level	30-min average	70
TVOC(μg/m³)			
Singapore		real time	3ppm
China-GB/T18883		real time	600
China HK-IAQC	Good level	8-hour average	600
	Excellent level	8-hour average	200
China-GB/T50325		real time	600
Temperature(℃)			
Singapore			22.5-25.5
China-GB/T18883	Summer		19-21
	Winter		16-24
American-ASHRAE	Summer	real time	23-26
	Winter		21-23
Europe-AQGE	Summer		22-28
	Winter		16-24
China HK-IAQC	Good	8-hour average	25.5
	Excellent	8-hour average	20-25.5
Relative Humidity（%）			
Singapore		real time	<70

China-GB/T18883	Summer		40-80
	Winter		30-60
Europe-AQGE	Summer		40-80
	Winter		30-60
Canada-EGR	Summer		30-80
	Winter		30-55
American-ASHRAE	Summer		50-60
	Winter		20-30
China HK-IAQC	Good	8-hour average	<70

Moreover, to obtain the average levels of residential indoor pollutants in a specific region, an appropriate number of dwellings are generally required in LTM. To study the risk of sick building syndrome in air-conditioned spaces, Cheung et al. [32] investigated the IEQ in 8 dwellings with different building areas, orientations and numbers of residents. Mentese et al. [33] carried out an LTM study in 121 dwellings located in three cities/towns in Turkey, to explore the relationship between respiratory diseases and indoor pollutants. Lim et al. [34] selected 25 apartments for a study of the relationship between IEQ and occupant health in energy-efficient dwellings. The purposes of these studies were different, but a large sample size would increase the financial cost of monitoring systems [35]. Therefore, it is meaningful to discuss the effect of the required number of dwellings (RND) on the results in a given region.



Journal Pre-proof

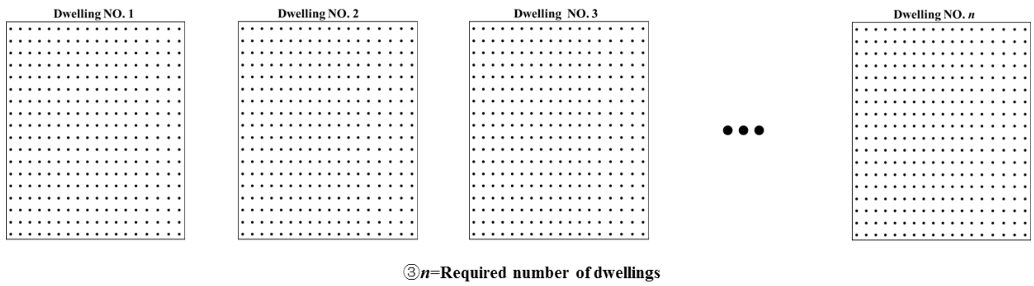


Figure 1 Sampling process of long-term monitoring data on residential indoor air quality

Hence, this study suggested that the IEQ data sample for LTM is determined by the sampling strategy, sampling frequency and RND, as shown in Fig. 1. The purpose of this work was to identify a suitable data sampling method for evaluation of IEQ, by means of the following steps: 1) study the periodic fluctuation characteristics of IEQ parameters in residential buildings, so as to propose sampling suggestions; 2) propose and compare continuous and discontinuous sampling strategies for different IEQ parameters; 3) design an algorithm based on the Pearson correlation coefficient to optimize the sampling frequency for various parameters, combining the experimental data for these parameters; and 4) calculate the RND for studying indoor air parameters according to seasonal and daily fluctuations, and for five cities in different building thermal zones in China. The method developed here will help researchers in the future to balance the required sample size and financial restrictions.

2. Methods

2.1 Data sources and long-term monitoring (LTM) method

To compare different LTM sampling strategies, this study employed IEQ-related field data collected from 13 residential buildings in Kunming, China, as shown in Fig.

Journal Pre-proof

Journal Pre-proof

2. The study focused on dwellings in the moderate climate zone of China. The buildings generally did not use heating and air conditioning systems, but only natural ventilation to regulate the indoor environment[36]. In order to analyze the impact of outdoor parameters on IEQ, we obtained the data of outdoor air parameters. The outdoor CO_2 concentration was measured by an IKAIR sensor that was located at the balcony of dwelling. The balcony was directly exposed to the outdoor environment and was not affected by indoor air conditions[8]. Outdoor PM2.5 concentrations were provided by the China National Environmental Monitoring Centre, which were collected from the nearest local meteorological and air quality station close to these dwellings[37]. The average outdoor PM2.5 concentrations of Kunming is 39 ug/m^3 during the monitoring period, which is one of the lowest among Chinese major cities[38]. By monitoring the outdoor and indoor CO_2 concentration, we noticed that the concentration difference is small[8]. Hence, the local outdoor pollution had less impact on IEQ. The main source of indoor pollutants was from indoors [39]. The case-study building was a high-rise structure, which is a common type of residential building in China [40]. The building dedication year has a greater impact on the intensity and decay rate of indoor pollution sources such as formaldehyde [41].

The long-term monitoring system for IEQ employed in this study was an integrated system with various gaseous sensors that monitored CO_2, $PM_{2.5}$, and formaldehyde concentrations, air temperature and relative humidity with a sampling period of one minute. In each dwelling, we placed one IKAIR sensor in the bedroom and one in the living room. The sensor is installed in the center of the monitored room, with a height of 1.5 m. According to previous published research [8], we can see that there a little difference in the monitoring results of various parameters between the bedrooms and living rooms of residences. Thus, we did not consider the effect of sampling locations in this study. The influence of sampling location on parameters is complicated, which is beyond the objective of this study. Therefore, this paper will not carry out in-depth discussion in this perspective. Table 2 lists the primary specifications of the sensors in this system. The data collected by the sensors are dynamically sent to a data center through the dwellings' Wi-Fi networks. Detailed information about the LTM system can be found in Liu et al. [37].

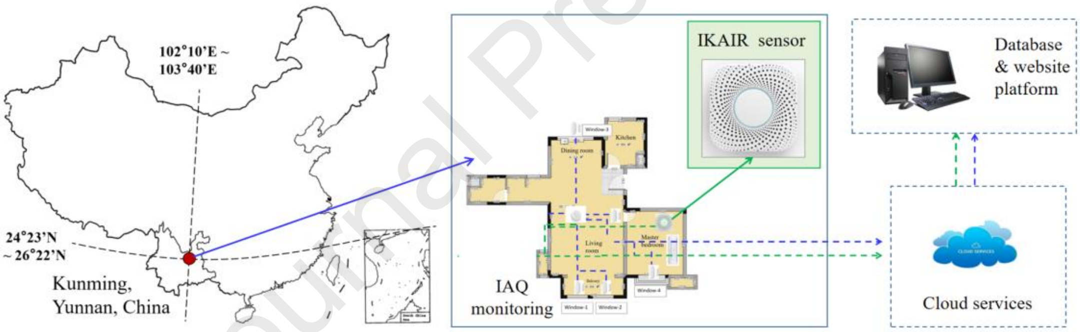


Figure 2 Long-term monitoring in residential buildings in Kunming

Before and after the monitoring, sensor calibration was carried out in a 1 m^3 integrated chamber, with a Dusttrak 8530, PB-RAE, temperature and humidity measuring instrument for the detection of PM2.5, formaldehyde, temperature and relative humidity, respectively. The Laskin atomization method [42] was employed to generate the particles and formaldehyde using standard sources, DEHS and trimetric formaldehyde. A constant temperature and humidity environmental cabin were used to measure the calibrated temperature of the sensor. The temperature measurement range

Journal Pre-proof

of the environmental chamber was -70–150 ℃ with an accuracy of 0.1 ℃, and the relative humidity measurement range was 20–98% with an accuracy of 0.1%. The calibration process exhibited a high regression coefficient of 0.99, indicating the high reliability of the measurement devices used in this study.

Table 2 Information about sensors used for long-term monitoring of IEQ

	Measurement principle	Measurement range	Accuracy
PM2.5 (ug/m³)	Laser scattering	1-1000	± 1
Formaldehyde (ug/m³)	Electrochemistry	0-5000	± 20
Temperature (°C)	Thermal resistor	-40-125	± 0.35
Relative humidity (%)	Humidity sensitive resistor	30-100	± 3
Carbon dioxide (ppm)	Carbon dioxide resistor	400-10000	± 30

2.2 Sampling strategies

This study proposed a continuous sampling strategy (CSS) and a discrete sampling strategy (DSS). Fig. 3 explains the difference between the two strategies by using a single day of formaldehyde measurements in a house as an example. The data are expressed in one-dimensional time-series graph and two-dimensional stacked graph. Fig. 3(a) shows the raw data obtained by measurement at a frequency of once per minute. In Fig. 3, the ordinate is the 24 hours within a day and the abscissa applies the time series of 60 minutes within each hour. Therefore, there are a total of $60\times24=1440$ data points as presented in Fig. 3(a). Varied data points can be extracted from the raw data by different sampling strategies, as shown in Fig. 3(b) and (c). For the DSS, one data point of the indoor parameters is acquired at each sampling time; thus, the data points collected by this method have no continuity in time. Meanwhile, as displayed in Fig.

Journal Pre-proof

3(b), formaldehyde concentration data points were sampled every three minutes. As shown in Table 1, the DSS exhibits significant differences across existing IEQ standards, and currently there is no conclusive reference for the appropriate DSS. For the CSS, data were sampled continuously. This continuity refers to the continuity of sampling of the raw data, so data are sampled at a certain sampling frequency to obtain multiple data points, as shown in Fig. 3(c). Therefore, the data represent IEQ conditions within a short period. As shown in Table 1, the CSS has commonly been employed in existing IEQ standards to determine the mean concentrations of indoor air pollutants, with duration of 8 hours, 24 hours or 1 year. Within these sampling durations, however, there are different frequencies for the data sampling. Nevertheless, the evaluation of IEQ was based on the average of the sampled data during the monitoring period.

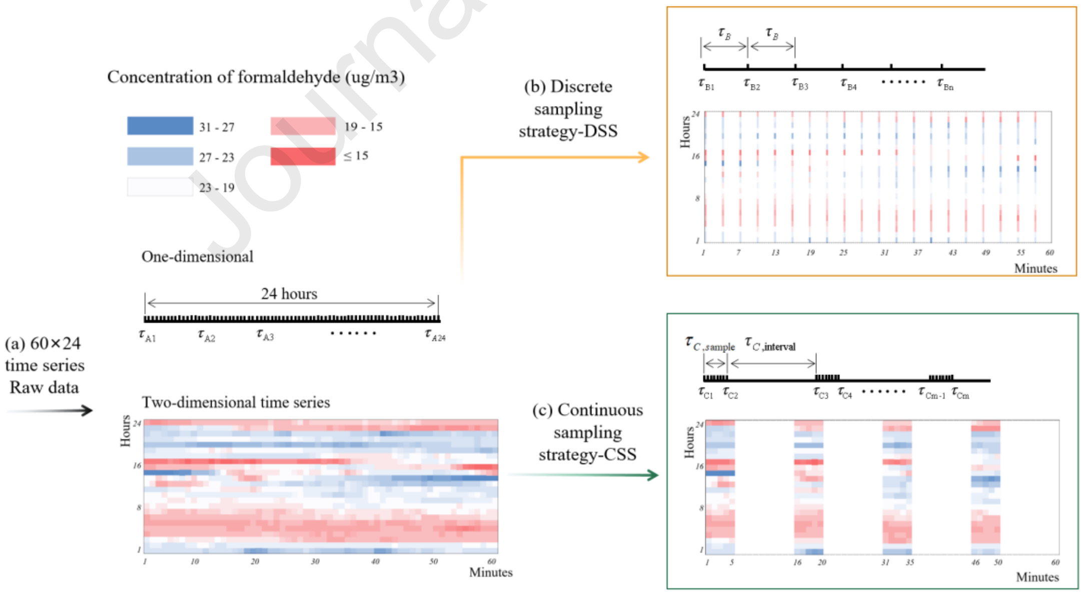


Figure 3 Principles of the DSS and CSS for the sampling of raw data

2.3 Data preprocessing method

Different indoor air parameters have different dimensions and orders of magnitude.

Journal Pre-proof

To evaluate and compare the sampling strategies for different IEQ parameters in parallel and to improve the credibility of the research results, it is necessary to preprocess the raw data obtained from LTM [43]. Commonly used data preprocessing procedures include missing value filling and data standardization processing [44]. Data standardization refers to scaling of the data to a small specific interval. After standardization, the data is transformed into a dimensionless pure value, so that comprehensive evaluation and analysis can be carried out. The z-score standardization method can be used. The processed data conform to the standard normal distribution; that is, the mean value is 0 and the standard deviation is 1, with the standardized form of data calculated by Eq. 1,

$$A_i(i=1,2,\ldots\ldots n)=\frac{A_i-\mu_A}{\sigma_A} \quad (1)$$

$$\mu_A=\frac{1}{n}\sum_{i=1}^{n}A_i$$

$$\sigma_A=\sqrt{\sum_{i=1}^{n}(A_i-\mu_A)^2}$$

where matrix $A_i(i=1,2,\cdots\cdots n)$ represents the raw data time series for LTM of IEQ parameters; $i=1,2,\cdots\cdots n$ represent the time scale dynamically measured IEQ parameters in the time series; μ_A and δ_A are the mean value and the standard deviation, respectively, of the raw data.

2.4 Evaluation method for various sampling strategies

The Pearson correlation coefficient (PCC) is one of the most widely used relationship measures [45] and is a statistical metric for the strength and direction of a linear relationship between two random variables[46]. Based on this attribute of PCC, we can use it to measure the correlation between the samples and the raw data to evaluate the performance of the two sampling strategies. In this study, there are two random variable matrices: raw data $A_i(i=1,2,\cdots\cdots n)$ and samples B_i $(i=1,2,\cdots\cdots n)$. The definition matrix $B_i(i=1,2,\cdots\cdots n)$ represents the sampled data obtained by sampling the raw data using the DSS or CSS. The PCC of matrices A_i and B_i is formally defined as the product of the covariance of two random variables divided by their standard deviations (which acts as a normalization factor). If each variable has n scalar observations, then the PCC of a certain IEQ parameter can be expressed by Eq. 2 [46],

$$R(A,B)=\frac{\operatorname{cov}(A,B)}{\delta_A\delta_B}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A-\mu_A}{\delta_A}\right)\left(\frac{B-\mu_B}{\delta_B}\right) \tag{2}$$

Here, μ_A and δ_A are the mean and standard deviation of $A_i(i=1,2,\ \ldots\ldots n)$, respectively, and μ_B and δ_B are the mean and the standard deviation of B_i $(i=1,2,\ \ldots\ldots n)$, respectively. The correlation coefficient $R(A, B)$ ranges from -1 to 1. The closer the absolute value of $R(A, B)$ to 1, the stronger the correlation between the two random variables.

Journal Pre-proof

2.5 Algorithm to determine sampling frequency

The quality of various sampling strategies can be assessed on the basis of the source data by means of correlation analysis. Since correlation analysis determines the optimal sampling frequency from the marginal value between correlation and frequency increase [47], it has been widely used to compare different data sampling strategies [45].

Journal Pre-proof

In this study, the correlation analysis was implemented in Python with the algorithm shown in Fig. 4.

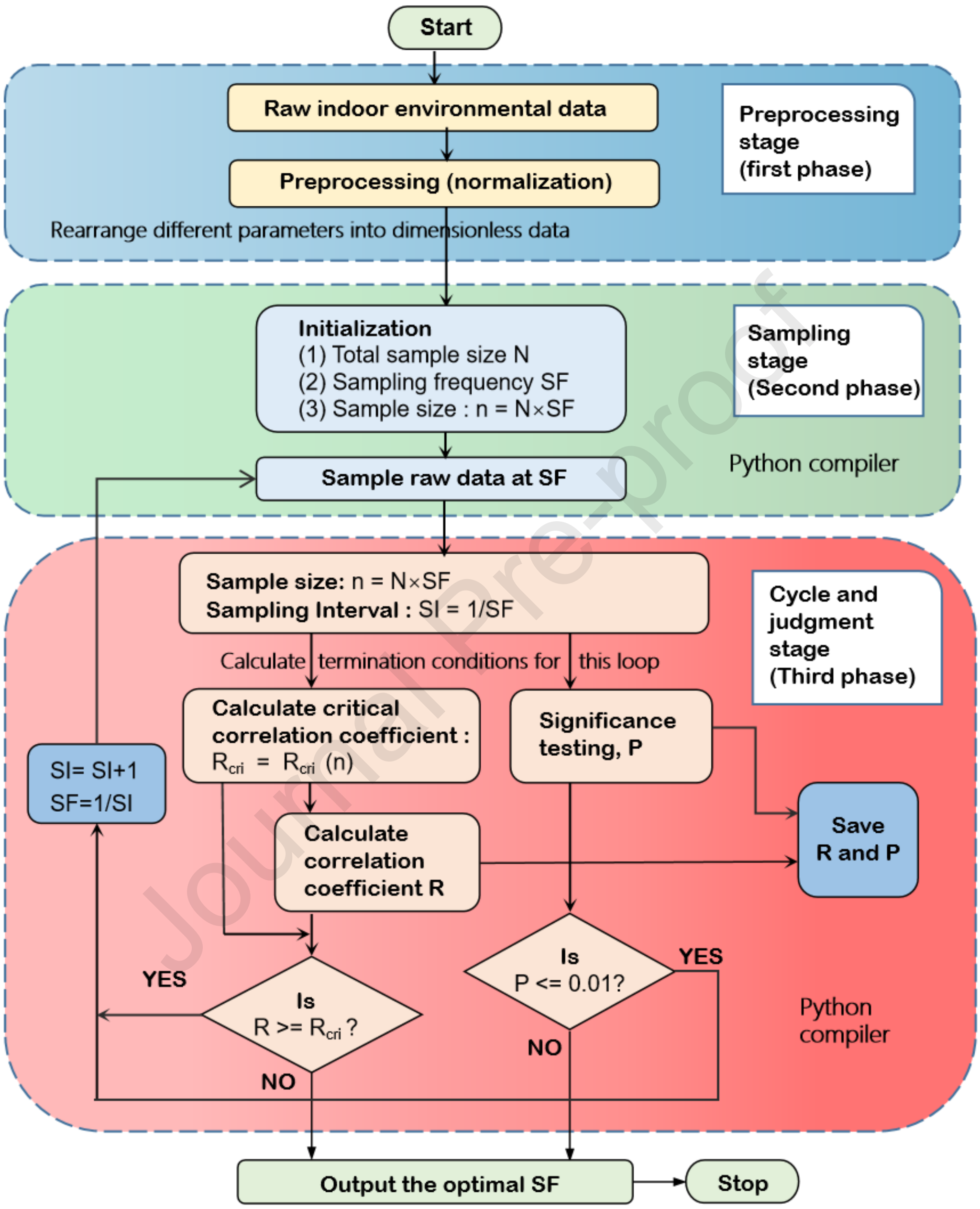


Figure 4 Algorithm for obtaining the optimal sampling frequency by correlation analysis

The algorithm reaches the optimal sampling frequency *SF* by continuously increasing the sampling interval *SI*. Each cycle uses the new *SI* as an intermediate variable to iterate the value of SF. The iteration uses PCC parameter *R* between the sampled data and the source data as the indicator variable. It updates the SI values until iteration stops when R is less than the critical Pearson correlation coefficient (CPCC) or passes the significance test P = 0.01. We looked up the CPCC value when $P = 0.01$, and the polynomial regression curve of the CPCC data was fitted in this study as shown in Fig. 4. The fitted regression curve accurately represents the condition of the loop termination in the algorithm.

$$\begin{aligned} R_{cri} =& 7.98\times10^{-19}\times n^{6} - 2.8\times10^{-15}\times n^{5} + 4.1\times10^{-12}\times n^{4} \\ & -3.06\times10^{-9}\times n^{3} + 1.26\times10^{-5}\times n^{2} - 0.003n + 0.45 \end{aligned} \tag{3}$$

In Eq. 3, R_{cri} and n represent the CPCC and the final sampling size, respectively. After testing, the polynomial regression curve represented the critical correlation coefficient with a degree of fit equal to 0.936. To avoid accidental error, at least 10 iterations were performed unless the R values converged. The effect of data dimensionality reduction was also analyzed after the optimum had been reached.

2.6 Method for determining RND

The sampled RND is an important factor in the time, cost and effort required for data collection. In the determination of the appropriate sampled RND n for a specific region, the choice of n from the total RND $n = 13$ is crucial in longitudinal studies to ensure that the results are representative for that region. Based on simple random sampling, n depends on the degree of overall difference, the allowable error, the confidence level and the adopted sampling method [48]. After these four items have been confirmed, n can be obtained by referring to the value that produces the minimum sampling error. In the simple random sampling method, n can be determined from Eq.

Journal Pre-proof

4,

$$n = \frac{z^2\sigma^2 N}{\Delta_{\bar{x}}^2 N + z^2\sigma^2} \tag{4}$$

where N is the total RND in the region; $\bar{x}$ is the overall average of the results for n dwellings; $\Delta_{\bar{x}}$ is the limited error of the $\bar{x}$; σ is the overall standard deviation; and z is the degree of probability, which is directly connected to the degree of confidence with a probability function of $F(z)$. In simple random sampling, $F(z)$ refers to the normal distribution function. When N is sufficiently large ($n/N \leq 5\%$), the RND = n can be determined by simplifying Eq. 4 into Eq. 5,

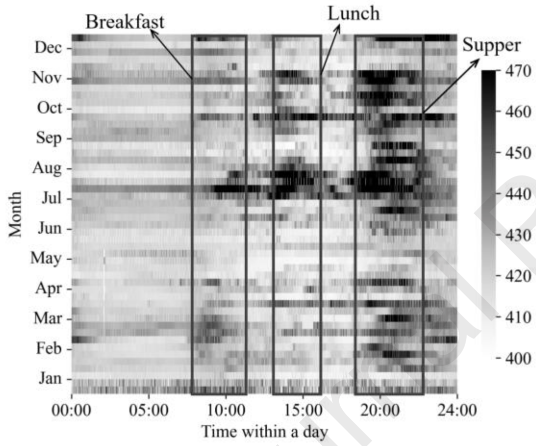
$$n = \frac{z^2 s^2}{\Delta_{\bar{x}}^2}, \text{ if } n/N \leq 5\% \tag{5}$$

3. Results and discussion

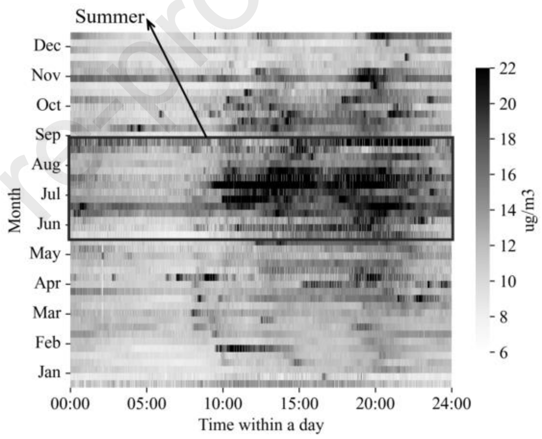
3.1 Heat map analysis of IEQ characteristics

Fig. 5 shows the indoor environment in the living room of one dwelling in the case-study building, with a monitoring period from 1st January to 12th December. In general, IEQ parameters would have certain changes in characteristics with time [49]. These characteristics are related to residents' behavior, environmental conditions and outdoor meteorological parameters. A heat map was produced for this dwelling, including annual indoor CO_2, formaldehyde, $PM_{2.5}$, TVOC, relative humidity, temperature. This map displays the concentration fluctuation characteristics of the measured data for each parameter during the overall monitoring period. The vertical axis represents the first week of the study. To display the characteristics of the parameters in different quarters more clearly, months were used as the vertical axis labels. In addition, to show the data characteristics of each parameter at different times during one day, the data for a 24-

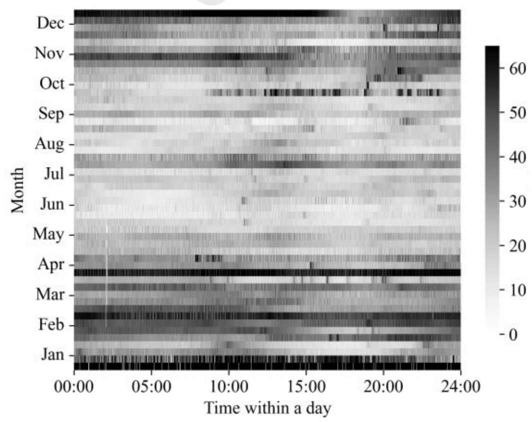
hour period was used to represent the indoor environmental quality for this week. The 24-hour data were the average of the continuous measurement data for the seven days of the week; therefore, they indicate the parameter characteristics for the week. The horizontal coordinate is in hours, from 0 to 24 hours. Analysis of the heat map allows the effects of residents' behavior, environmental conditions and outdoor parameters on the long-term measurement data to be identified. Next, the sample duration can be calculated, the sampling frequency that best represents the indoor air quality can be determined, and the appropriate sampling strategy can be proposed.



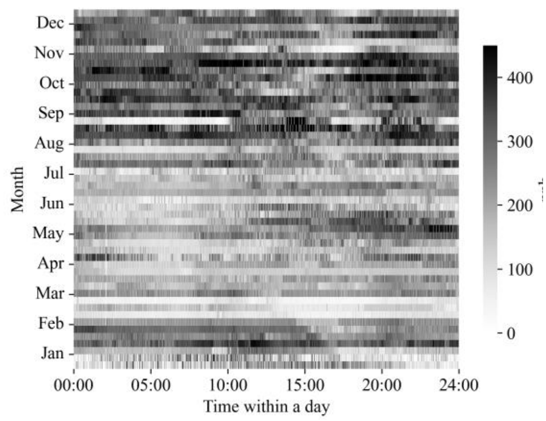
(a) Carbon dioxide



(b) Formaldehyde

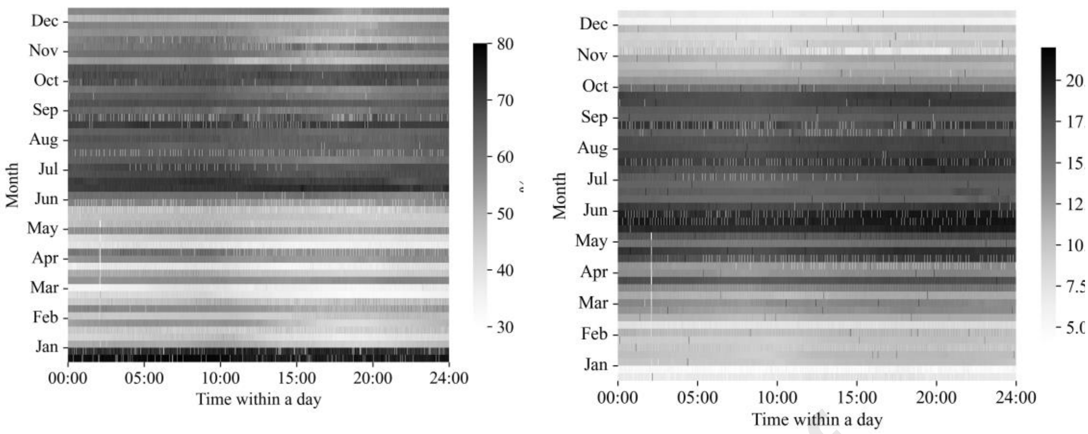


© PM2.5



(d) TVOC

Journal Pre-proof



(e) Relative humidity (f) Temperature

Figure 5 Heat map of measured carbon dioxide, formaldehyde, PM2.5, TVOC, relative humidity and air temperature for the 2017–2018 period in a dwelling in Kunming, China

According to previous research[8, 50], indoor CO_2 emission sources mainly involve human breathing and fuel combustion in kitchen. Hence, it is generally believed that these two sources of CO_2 are present only when the residents are at home [51]. CO_2 is an indicator of the general level of air pollution related to the presence of humans indoors and can therefore reflect the level of human exposure [52]. It can be seen in Fig. 5(a) that the concentration of CO_2 changed with time. In the course of a day, both the value and the gradient of the concentration varied greatly. As shown in Fig. 5(a), the CO_2 concentration value was much higher during the periods from 11:00 to 15:00 and from 19:00 to 21:00 than during the rest of the day. The value between 7:00 and 10:00 was also higher than the daily average. The value was significantly lower during the period from 0:00 to 7:00. From this measurement, it can be concluded that the outdoor CO_2 concentration was quite stable at about 400 ppm, so the change in CO_2 concentration indoors was driven mainly by indoor sources. The peak CO_2

concentration was maintained for the most part in the 460–480 ppm range, which meets the requirements of the Chinese IEQ standard (GB 18883). According to the peak concentration of CO_2 on the heat map, cooking behavior increased with the indoor CO_2 concentration during meal times, especially at lunch and dinner. Therefore, attention should be paid to sampling periods with high CO_2 concentrations, especially during cooking times. Large variations in the average value and the gradient of the CO_2 concentration can reflect the occupants' behavior, e.g., the fact that the residents are at home. Hence, obtaining more detailed information requires high-frequency sampling.

Meanwhile, changes in the indoor formaldehyde concentration throughout the year are shown in Fig. 5(b). The peak concentration of formaldehyde within a day mainly appeared at daytime from 10:00 to 21:00, and the concentration of formaldehyde is relatively lower at night. Moreover, the gradient of formaldehyde in different seasons within a year is also large. Among them, the peak value of formaldehyde mainly appeared in summer (June, July, and August). As shown in Fig. 5(b) and (f), it can be deduced that the change of temperature will cause the change of formaldehyde concentration. The results is similar to another research, which reported the formaldehyde emission from decoration and furniture materials surfaces is closely correlated with air temperature [53]. In some cases, high formaldehyde concentrations are difficult to predict, so the best observation times for formaldehyde concentration can be determined by monitoring the periods with high air temperature.

As shown in Fig. 5(c) and (d), PM2.5 and TVOC concentrations did not fluctuate significantly in the course of a day. However, the variations from one week to another

Journal Pre-proof

were large. In Fig. 5(c), it can be seen that the PM2.5 concentration had relatively high values from November to April, when the outdoor PM2.5 concentration was high. Fig. 5(d) indicates that the concentration of indoor TVOCs was relatively high in January, February, and from August to December, but low in other months. It is generally believed that indoor TVOCs are emitted mainly by building and decoration materials [54] and are independent of air temperature and humidity; therefore, the main factor in the concentration is the air change rate of the dwelling. In autumn and winter, occupants tend to close windows and doors so that the air exchange rate is lower than in other seasons, which may have caused the increase in TVOC concentration. The indoor PM2.5 concentration exhibited a similar trend to that of the TVOCs, as can be seen in Fig. 5(c). The PM2.5 and TVOC concentrations did not vary greatly in the course of a day, possibly because the occupants were nonsmokers or did not smoke inside the dwellings [55]. Above all, when detecting PM2.5 and TVOC concentrations, it is recommended that the effects of the season and the fluctuations within months and seasons be taken into account.

It should be noted that temperature and humidity vary greatly from season to season. As shown in Fig. 5(e), the relative humidity in July, August, September and October is high (65–75%), and the relative humidity values from February to June and in November and December are low. It can be seen in Fig. 5(f) that the distribution of the temperature heat map is also related only to the month. May through October are the peak months, with a temperature range of 16—22 °C, and the temperature in the other months is lower, ranging from 5°C to 13 °C. The dwelling in this study was located in the moderate climate zone of China, with small annual temperature differences and discrete dry and wet seasons. Summer and autumn in this region are rainy, and the outdoor humidity is relatively high. Therefore, on the heat map, the measured relative humidity inside residential buildings in summer and autumn is relatively high, and the temperature in summer is also relatively high. However, the temperature and humidity did not fluctuate significantly in the course of a day. Therefore, it is recommended that the seasonal characteristics be taken into account when sampling. Since a numerical gradient was observed only from week to week, the sampling of temperature and relative humidity can be performed at a weekly frequency.

Fig. 5 presents all the data collected for the dwelling throughout the year. The figure also shows detailed characteristics as follows: 1) according to periodic variation characteristics of CO_2, formaldehyde, PM2.5, TVOC, relative humidity and air temperature, the cyclical behavior of the occupants in residential dwellings can be preliminarily observed; and 2) there are large differences between IEQ parameters, so the sampling strategy should be formulated according to the time fluctuation characteristics of the parameters.

3.2 Evaluations of different sampling strategies

In the study, we evaluated two sampling strategies: the discrete sampling strategy (DSS) and the continuous sampling strategy (CSS). With the use of continuously measured data for six parameters, PM2.5, formaldehyde, TVOC, CO_2, temperature and relative humidity, in a residence in Kunming throughout the year, sampling with different measurement intervals was performed by means of the DSS and CSS.

Journal Pre-proof

The sampling period was set to hours, days and weeks. And in each sampling period, i samples were chosen from the raw data. The sampling method is abbreviated as Sampling Strategies (SS) - Sampling period-i. Different sampling strategies will extract the same amount of data during the same sampling period. Therefore, the final sample sizes obtained by the two sampling strategies were the same.

Statistical analysis could easily and intuitively process an array of time series data. it plays an important role in comparing the representativeness of data obtained under different sampling strategies. In this study, frequency distribution histogram analysis and descriptive statistics analysis were performed on different parameters. Considering that Pearson Correlation Coefficient (PCC) can be used to calculate the correlation between two arrays, by comparing the PCC calculated by raw data and the sampled data, we can obtain the optimal sampling strategy.

3.2.1 Effects of different sampling strategies on data distribution

Using the CSS and DSS, we sampled the annual data for PM2.5 and CO_2 in the case-study dwelling. In order to compare the distribution of the sampled data with the raw data, a histogram of the frequency distribution under different sampling frequencies was plotted, as shown in Fig. 6. The IEQ data were sampled at frequencies of hours, days and weeks.

Journal Pre-proof

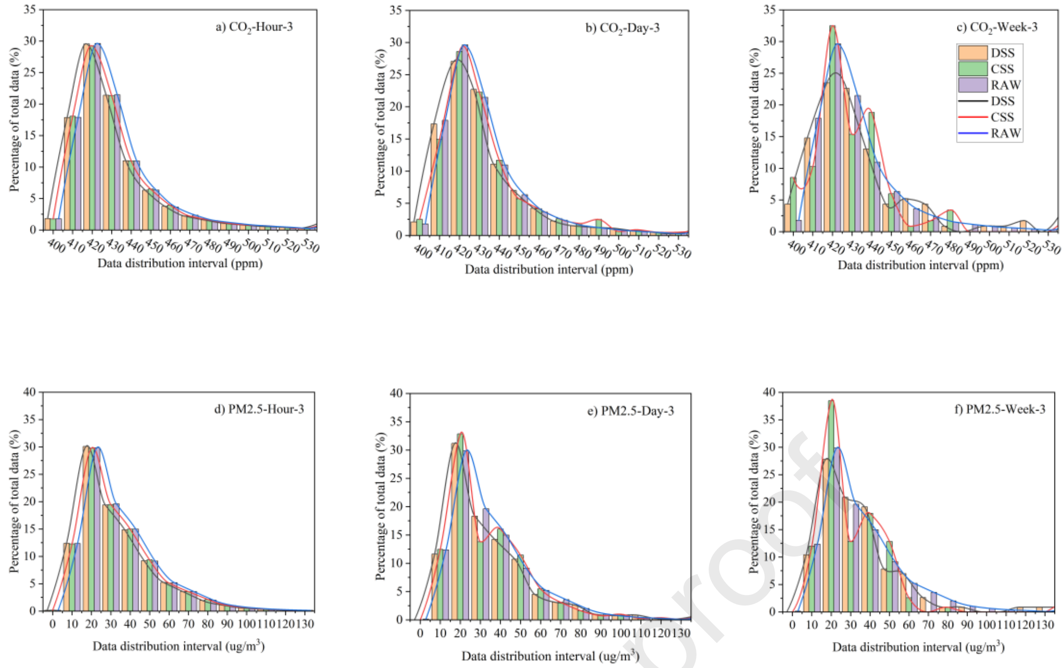


Figure 6 Density of distribution of PM2.5 and CO_2 concentrations at various frequencies under the DSS and CSS in a single dwelling in Kunming, China; histograms (a), (b) and (c) refer to CO_2, and (d), (e) and (f) refer to PM2.5

The density distribution exhibited a positive trend with the sampling frequency, as shown in Fig. 6. According to Fig. 6(a) and (d), when CO_2 and PM2.5 were sampled at a frequency of once every 3 hours, the density distribution under both the CSS and DSS was close to that of the data source. When the sampling frequency was once every 3 hours, the density distributions under the CSS and DSS have little difference. With a sampling frequency of 3 days as shown in Fig. 6(b) an©(e), the density distribution under the CSS differed greatly from that under the DSS. When the sampling frequency was reduced to once every three weeks as shown in Fig. 6(c) and (f), the distributions of the DSS and CSS differed more strongly. Moreover, the DSS generally yielded a better comparison with the raw data, especially for the peak values. The data sampled

Journal Pre-proof

under the CSS fluctuated less, with a more scattered distribution. The density distribution in general prohibits a lognormal distribution. From the above, it can be concluded that the sampling frequency should be on the order of hours or days, but no less frequently than week for CO_2 and PM2.5.

3.2.2 Impact of different sampling strategies on descriptive statistics

The study investigated not only the average value but also the standard deviation, coefficient of variation, partial peak, and kurtosis value. The average value indicates the overall level of the sample, and the standard deviation and coefficient of variation reflect the degree of dispersion of the data. The skewness and kurtosis value measure the degree of skewness and flatness of the distribution, respectively. The maximum and minimum values indicate the data range. It is recommended that a sampling strategy be selected that has closer agreement with the raw data as well as fewer data points.

Journal Pre-proof

Table 3 Comparison of DSS and CSS by descriptive statistics for PM2.5 and CO_2

Sampling strategy	Mean	Standard deviation	Correlation of variance	Partial peak	Kurtosis	Minimum	Maximum	Data points
PM2.5								
Raw data	30.4	24.33	0.8	3.4	28.6	0	852	384694
DSS-Hour-3	30.43	24.55	0.81	3.41	25.5	0	489	19235
CSS-Hour-3	29.8	23.56	0.79	2.73	13.61	0	243	19236
DSS-Day-3	30.86	24.39	0.79	2.73	13.43	1	229	802
CSS-Day-3	30.29	23.74	0.78	2.74	13.58	1	213	802
DSS-Week-3	32.21	25.58	0.79	2.92	13.4	2	189	115

Journal Pre-proof

CSS-Week-3	29.57	31.85	1.08	4.39	22.49	1	213	117
Carbon dioxide								
Raw data	427.81	27.59	0.06	5.59	133.26	400	2087	384694
DSS-Hour-3	427.9	29.42	0.07	11.9	526.91	400	2070	19236
CSS-Hour-3	427.91	27.55	0.06	4.79	63.57	400	1121	19235
DSS-Day-3	428.29	27.23	0.06	4.53	45.51	400	800	802
CSS-Day-3	428.79	24.71	0.06	1.91	4.26	400	548	804
DSS-Week-3	434.13	44.62	0.1	5.35	39.8	400	800	115
CSS-Week-3	426.57	24.35	0.06	2.48	8.65	400	538	117

In previous studies, the maximum, minimum and average values have been used as evaluation criteria. The mean values differed by less than 5.9% between the DSS and CSS. The standard deviation and coefficient of variation for PM2.5 varied greatly with sampling frequency under the DSS. The accuracy of the monitoring instrument of CO_2 is ±30 ppm, yet the standard deviation of the data is lower than the accuracy. Therefore, we do not study the average of CO_2. According to a comprehensive analysis of peak and kurtosis values for CO_2 and PM2.5, the DSS exhibited closer agreement with the raw data. The maximum value also indicates a closer agreement under the DSS than the CSS. The number of data points is another important factor in the choice of sampling frequency.

We noticed that the peak concentration of CO_2 is an important IEQ parameter. Since if the maximum of CO_2 is high, more fresh air needed to be supplied to the house

Journal Pre-proof

to improve the IAQ. According to Table 1, the threshold peak concentration of CO2 in China is 1000 ppm, while in WHO is 900 ppm. Table 3 shows the maximum obtained by DSS is much higher than CSS and is closer to the raw data. In other words, the DSS data are more likely to indicate that the house needs fresh air. Thereby, it is recommended to use DSS-Hour-3 for the sampling of CO_2.

In short, although the statistical measures of mean, standard deviation, and coefficient of variation are not regular, analysis of the data for other IEQ parameters is still needed. However, the descriptive statistical characteristics of the DSS samples are considered to be closer than CSS in terms of partial peaks, kurtosis values and sample ranges.

3.2.3 Correlation analysis of different sampling strategies

The DSS and CSS were used to sample PM2.5, temperature(T), relative humidity(RH), formaldehyde, CO_2 and TVOCs in a residence in Kunming throughout the year, and the Pearson correlation coefficient(PCC) between the sample and the raw data was compared for the two strategies. The PCC reflects the correlation between samples. The calculation of the correlation coefficient requires that the variables have the same sample size. Because the data distributions of DSS-Week-*i* and CSS-Week-*i* differ greatly from the raw data according to the conclusion of section 3.2.1, this comparison only involves the sampling strategies of DSS(CSS)-Day-*i* and CSS(CSS)-Hour-*i*. The limits set by China-GB/T18883, China-JGJ/T309, and China-GB/T50325 for residential PM2.5, formaldehyde and CO_2 are all daily averages. Therefore, we calculated the daily average for the data obtained with different sampling frequencies and different sampling strategies to unify the sample size. We were then able to analyze the correlation coefficient between the daily average value and the raw data.

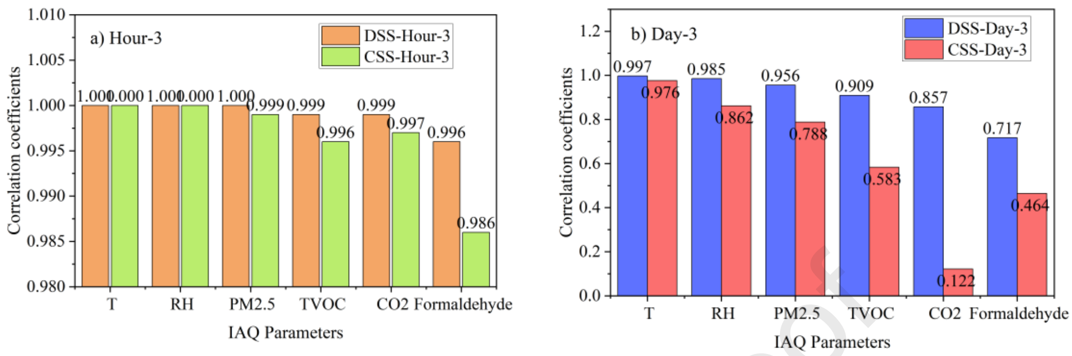


Figure 7 Comparison of DSS and CSS in terms of correlation coefficients for six IEQ parameters: (a) DSS-Hour-3, CSS-Hour-3; (b) DSS-Day-3, CSS-Day-3

Fig. 7 displays the results for the DSS and CSS with sampling frequency of once every three hours or once every three days, referred to as Hour-3 and Day-3, respectively. The Pearson correlation coefficient of the DSS, namely, DSS-Hour-3 and DSS-Day-3, for the six IEQ parameters indicates a closer relationship with the raw data than does the PCC of the CSS. The linear relationship between the sampled data under the DSS and the raw data is obvious. Therefore, the samples obtained under the DSS more accurately reflect the characteristics of IEQ.

3.3 Optimal sampling frequency under DSS

According to the descriptive statistics and correlation coefficients for the two strategies, the sampled data obtained under the DSS are superior to the CSS data in reflecting the characteristics of the raw data. As discussed in Section 3.1, the values of the IEQ parameters fluctuated periodically with the days and weeks. In order to understand the differences among various parameters, it is necessary to further study the sampling frequency and determine the characteristics of the data.

Journal Pre-proof

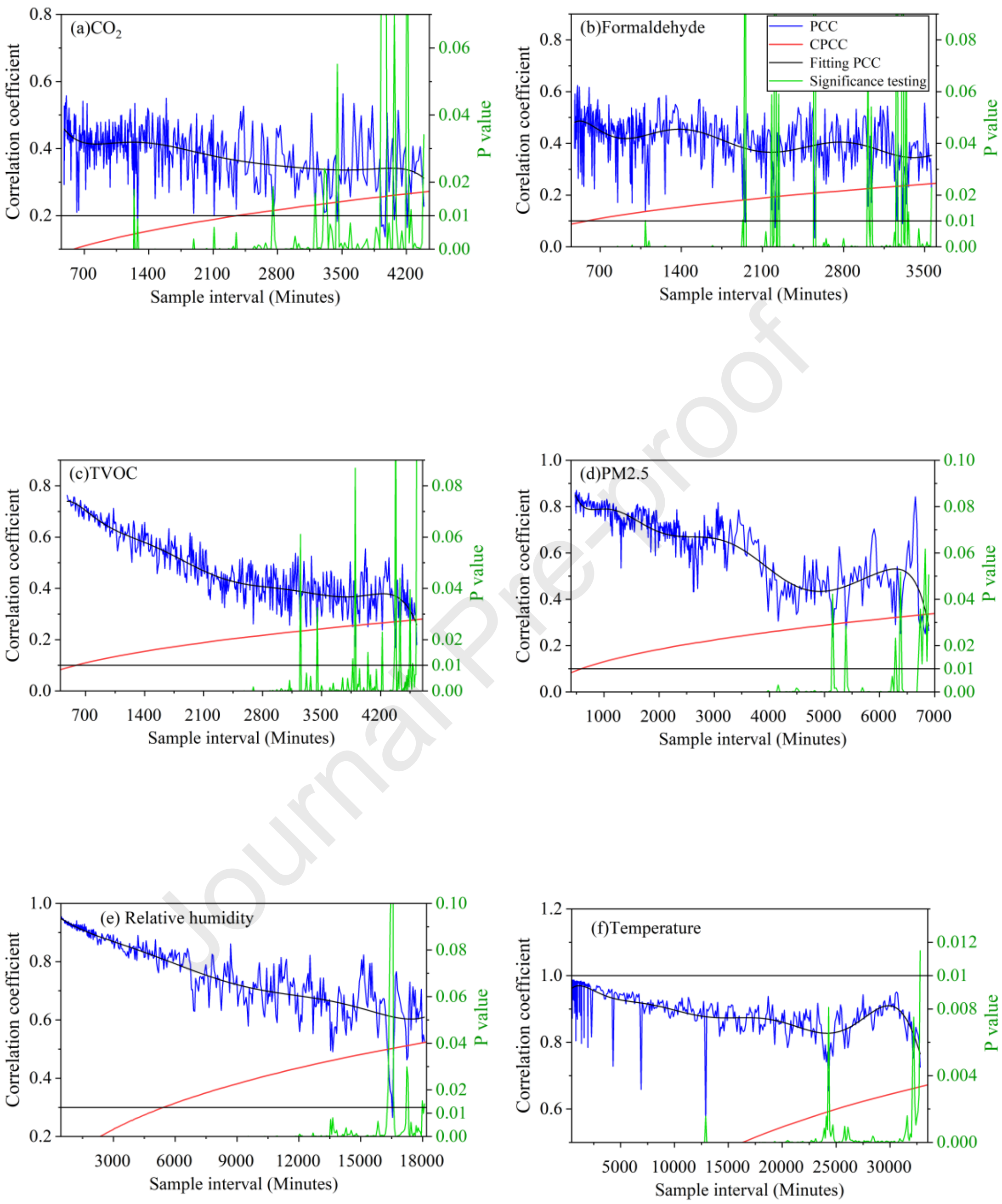


Figure 8 PCC and significance test for sampling frequencies under DSS for dix indoor air parameters:(a)Carbon dioxide, (b)Formaldehyde, (c)TVOC, (d)PM2.5, (e)Relative humidity,(f)Temperature. Legends are indicated in (b):CPCC means critical Pearson correlation coefficient; PCC, CPCC refers to the coordinates on the left; Significance test results refer to the coordinates on the right

Journal Pre-proof

The calculated results for PCC under the DSS are shown in Fig. 8, where the sample interval was equal to the inverse of the sampling frequency. As the sampling interval increased, the PCC of the sample gradually decreased, but the CPCC gradually increased. The fitted curves converged at the point where the sampling interval reached its optimum. At the same time, we took into account the significance of the results. When the *P* value was less than 0.01, we considered the test to be unqualified and the sample data to no longer be correlated with the raw data. When the significance test failed, the two curves might not intersect, and the iteration stopped.

Fig. 8(a) shows the PCC and CPCC data for CO_2 and the significance test results with the change in sampling interval. The optimal sampling interval obtained by the algorithm was 4390 minutes; that is, the sampling period was 3.04 days, which accounts for only a single data point every three days. Regardless of accidental errors, the correlation coefficients at certain frequencies also reached a critical value (intersection point). After the determination of correlation coefficients, the significance test was performed. According to the test result, the null hypothesis that the sampled data is not correlated with the raw data should be rejected. Therefore, the optimal sampling frequency of CO_2 is at least once every 3 days.

The calculated sampling frequency for formaldehyde in Fig. 8(b) is once every 2.47 days. Thus, the formaldehyde in this dwelling should be sampled at least once every 2 days. According to the calculation result for TVOCs in Fig. 8(c), the sampling interval is 4630 minutes and the sampling frequency is once every 3.2 days, which means that the optimal sampling frequency for TVOCs is at least once every 3 days. Meanwhile, the sampling interval for PM2.5 in Fig. 8(c) is 6890 min with an optimal sampling frequency of at least once every 4 days. The sampling intervals for relative humidity and temperature in Fig. 8(e) and (f) are 18090 and 32800 minutes, respectively, with optimal sampling frequencies of at least once every 12 and 22 days.

In many studies, the IEQ parameters have all been monitored with the same frequency by integrated sensors that are similar to the ones in this study. In cases in which six IEQ parameters are measured simultaneously with the same sampling frequency, the frequency that is selected should be the one that is the lowest among all the parameters. In the dwelling used in the present study, the lowest frequency was that for formaldehyde. The PCC and significance test are therefore highly recommended for the determination of optimal sampling frequency, not only for individual sensors but also when integrated sensors are used.

3.4 RND for seasonal and daily average data

For studies of different durations, the number of samples to be monitored is often different. In this section, residential sample sizes required for the study of indoor air parameters are proposed in accordance with seasonal and daily fluctuations of these parameters in different dwellings [13, 37, 56]. In terms of seasonal averages, different numbers of dwellings are required for each season. As shown in Fig 9, the seasonal RND ranges from 10 to 100 for PM2.5, from 15 to 40 for formaldehyde, from 2 to 17 for temperature, and from 1 to 3 for relative humidity. Therefore, PM2.5 requires the largest RND for long-term monitoring.

Journal Pre-proof

Journal Pre-proof

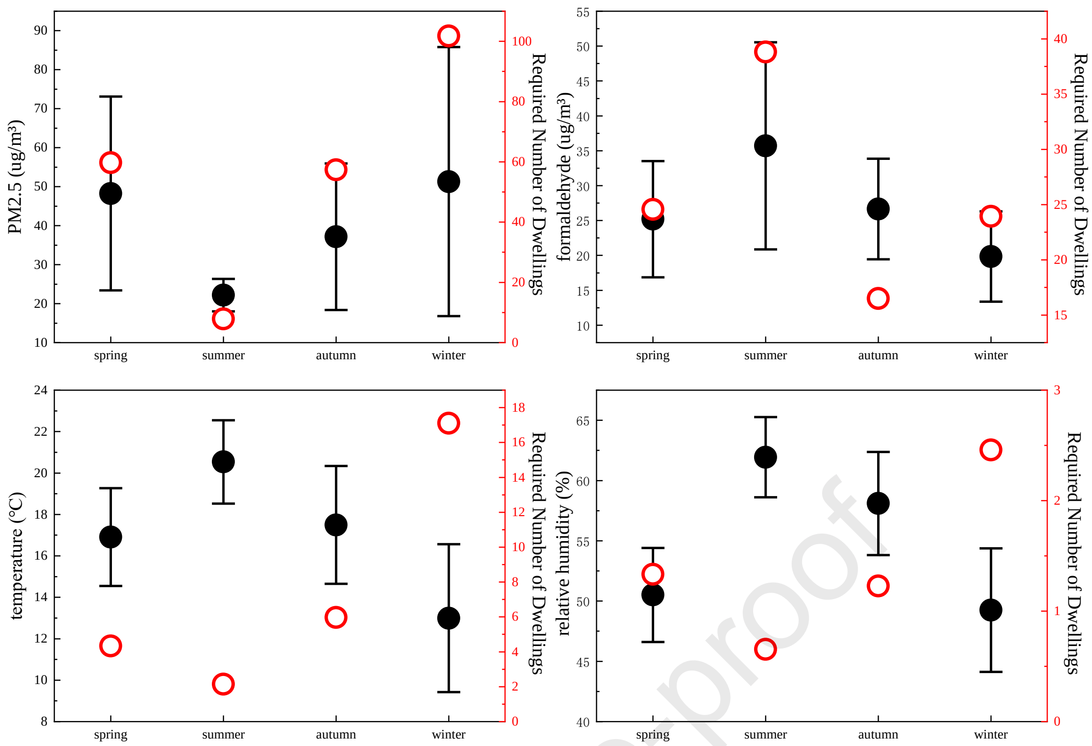


Figure 9 RND for determining seasonal averages for indoor PM2.5, formaldehyde, air temperature and relative humidity in Kunming, where black dots and red circles represent the change of averages and RNDs, respectively

The RND for PM2.5, temperature, and relative humidity exhibited similar trends with seasonal changes. The seasonal RND is the largest in winter and the smallest in summer. Meanwhile, it is equivalent in spring and autumn. The seasonal RND is affected neither by the sample mean and the individual variance of the sample, nor by the ratio of the sample variance to the sample mean.

Formaldehyde is distinct from PM2.5, temperature, and relative humidity. The change in seasonal RND with the season for PM2.5, temperature, and relative humidity in the 13 dwellings was close to each other. Formaldehyde has the largest RND in summer because of a large difference among dwellings in the strength of the formaldehyde emission source. The variations in temperature and humidity in summer

Journal Pre-proof

have a significant effect on the release of formaldehyde from decorative products. This finding agreed well with those of related studies [57-59], which reported formaldehyde emissions at high levels when temperature and humidity were high.

Fig. 10 displays a distribution histogram of the daily RND for PM2.5, formaldehyde, temperature and relative humidity [13, 60]. The RND satisfies the lognormal distribution throughout the year. The mean daily RND for PM2.5, formaldehyde, temperature and relative humidity are 11, 39, 3 and 2, respectively, and the variance of the RND is 0.74, 0.67, 0.51 and 0.66. The annual, seasonal, monthly or daily RND for temperature and humidity is less than the RND for PM2.5 and formaldehyde. Indoor temperature and relative humidity varied only slightly among dwellings because they were affected mainly by the outdoor climate conditions. In contrast, indoor PM2.5 and formaldehyde varied greatly because the strength of indoor pollutant sources differed from one dwelling to another.

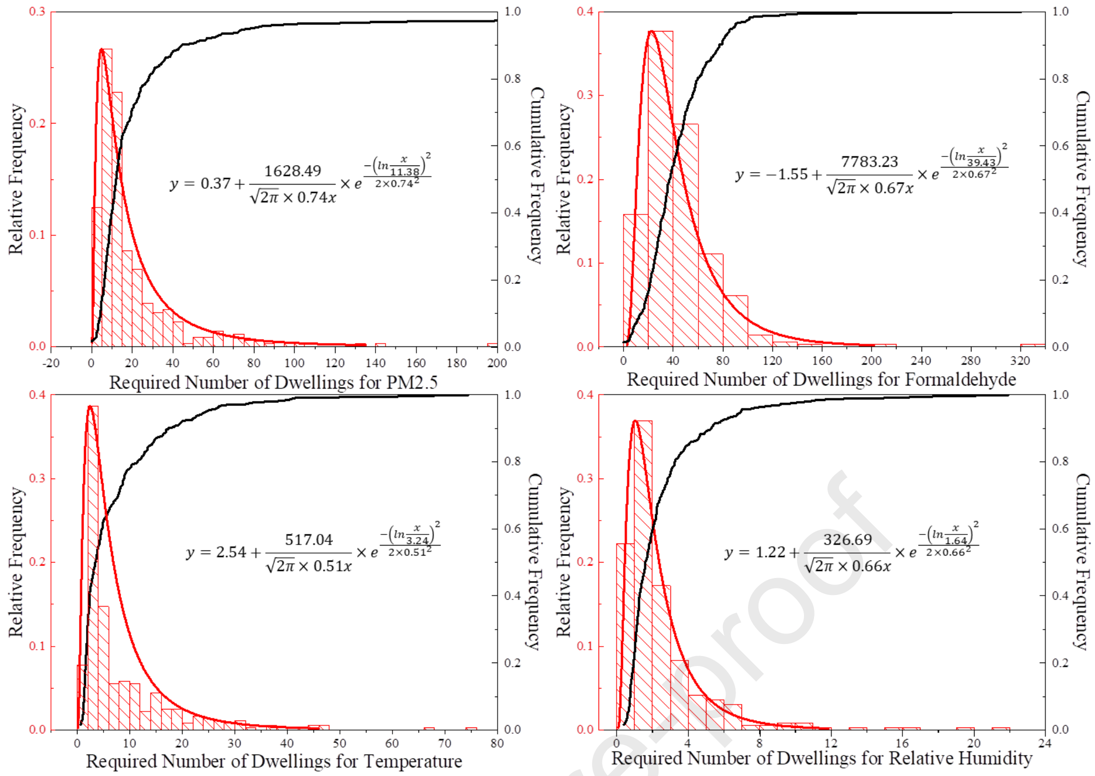


Figure 10 RND for daily average of indoor PM2.5, formaldehyde, air temperature and relative humidity. Fitting formulas are logarithmic normal distributions of relative frequency

Based on the cumulative frequency of the RND (see Table 4), when the confidence level is 90%, the RND for PM2.5, formaldehyde, temperature and relative humidity are 45, 77, 20, and 5, respectively. When the confidence is 95%, the RND for PM2.5, formaldehyde, temperature and relative humidity are 78, 87, 28, and 7, respectively. Therefore, the greater the confidence level, the larger the RND. Hence, when calculating RND, it is necessary to take into consideration either the confidence level or the number of data points.

Table 4 Relationship between RND and degree of confidence

Confidence level	PM2.5 (ug/m³)	Formaldehyde （ug/m³）	Temperature (℃)	Relative humidity (%)

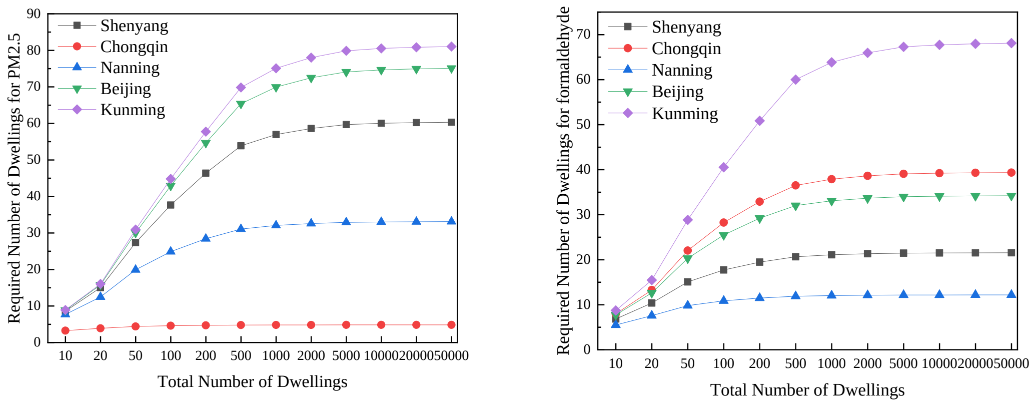
Journal Pre-proof

90%	45	77	20	5
95%	78	87	28	7

3.5 RND in various thermal zones in China

It can be seen in Fig. 11 that different thermal zones have different RND for studying the levels of PM2.5 and formaldehyde in a region. The RND is affected mainly by the coefficient of variation of the study parameters. The larger the coefficient of variation, the greater the difference in the concentration of pollutants between dwellings, and the larger the RND. Based on the current monitoring data, the difference in formaldehyde between Chongqing and Nanning is greater than that for PM2.5, while the difference in formaldehyde among Kunming, Beijing, and Shenyang is less than that for PM2.5. At the same time, a larger coefficient of variation leads to a larger RND. In Kunming, Beijing and Shenyang, the PM2.5 varies greatly among dwellings, while in Chongqing and Nanning, the difference in PM2.5 is small. For PM2.5, the RND is 60 in Kunming, while the RND in Beijing and Shenyang is less than 35, and in Chongqing and Nanning it is less than 5. For formaldehyde, the RND in each city is less than 20.

Journal Pre-proof



(a) (b)

Figure 11 Relationship between RND and total RND *N* for indoor PM2.5 (a) and formaldehyde (b) in Shenyang, Chongqing, Nanning, Beijing and Kunming

Fig. 11 shows the relationship between the total RND *N* and RND for PM2.5 and formaldehyde in five large cities in China. It can be seen that as the total RND increases, the RND increases as well, but when the total RND exceeds a certain value, the RND no longer increases. In this context, the threshold is called the critical total RND. If the number of dwellings in the region is estimated to be more than the critical total RND, *N* is considered to be equal to the critical total RND. In the five cities of Kunming, Beijing, Nanjing, Chongqing and Shenyang, the critical total RND *N* was approximately 5000 for the calculation of RND with $\Delta_{\bar{x}}=0.1\bar{x}$ and $F(t)$ = 87%.

4. Conclusion

This study has proposed a series of systematic methods for the sampling of indoor environmental parameters in urban residential dwellings. This paper starts with the sampling strategies and sampling frequency for various indoor environmental parameters in a single residence, and finally proposes the total required number of

Journal Pre-proof

dwellings for cities in different thermal zones of China. Compared with sampling methods that correspond to the thresholds of various standards, the approach in this study is more systematic. In order to focus on the periodic characteristics of various parameter fluctuations, this study used statistical methods, thereby enhancing the representativeness and credibility of the samples. The relevant findings can be summarized as follows.

1) A heat map revealed obvious daily and weekly fluctuations in CO_2, TVOCs, formaldehyde, PM2.5, air temperature and relative humidity. The CO_2 concentration reaches a peak during meal times. High concentrations of formaldehyde occur in periods with high temperature. PM2.5 exhibits a high concentration in autumn and winter. The distributions of other parameters display various seasonal fluctuations.

2) Because of the variations in environmental parameters inside residential dwellings, two sampling strategies, the continuous sampling strategy (CSS) and the discrete sampling strategy (DSS), were compared by means of descriptive statistical analysis. In general, the DSS performs better than the CSS in terms of accuracy.

3) Using the DSS as the sampling strategy, this study proposed an algorithm to calculate the optimal sampling frequencies for different indoor environmental parameters. We found that the optimal sampling frequencies for the concentration of TVOCs, PM2.5, CO_2, formaldehyde, relative humidity and temperature are 3 days, 4 days, 3 days, 2 days, 12 days and 22 days, respectively. The algorithm can effectively extract the periodic fluctuation characteristics of different indoor environmental parameters, thus providing more representative indoor environmental quality data and

Journal Pre-proof

effectively reducing sampling costs.

4) A method was proposed for determining the required number of dwellings (RND) in different thermal zones. Based on simple random sampling, the RND for studying different indoor environmental parameters was calculated. The results of the study can be confirmed with the first and third conclusions; that is, the degree of a parameter's fluctuation determines the sample size that is needed to accurately reflect the characteristics of the data. The required number of dwellings depends on the coefficient of variation of the sampled data. PM2.5 and formaldehyde had greater RND values than did temperature and humidity. The RND satisfies the lognormal distribution throughout the year.

However, it must be admitted that this study has certain limitations. First, the methodology of this study is mainly based on statistical principles and cannot be used trace the sources of pollutants. It can only make certain assumptions based on the fluctuations of parameters. Second, the raw data is limited to that from residential dwellings in Kunming, China. Nevertheless, the series of sampling strategies and algorithms proposed in this paper are applicable to other types of sampling processes and to residential environment research in other regions. Since the proposed algorithm can effectively extract the periodic fluctuation characteristics of different indoor environmental parameters and reduce sampling costs, we expect that the method will provide valuable assistance to future IEQ long-term monitoring studies.

Journal Pre-proof

5. Acknowledgements

The authors are very thankful for the support of students and colleagues at Tianjin University. The authors are also thankful for the financial support provided by the Natural Science Foundation of Tianjin (Grant No. 19JCQNJC07100).

References

[1] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, W.H.J.J.E.A.E.E. Engelmann, The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants, 11(3) (2000) 231-252.
[2] H.-J. Oh, N.-N. Jeong, J.-R. Sohn, J. Kim, Personal exposure to indoor aerosols as actual concern: Perceived indoor and outdoor air quality, and health performances, Building and Environment 165 (2019).
[3] S. Mentese, N.A. Mirici, M.T. Otkun, C. Bakar, E. Palaz, D. Tasdibi, S. Cevizci, O. Cotuker, Association between respiratory health and indoor air pollution exposure in Canakkale, Turkey, Building and Environment 93 (2015) 72-83.
[4] F. Wu, D. Jacobs, C. Mitchell, D. Miller, M.H.J.E.H.P. Karol, Improving Indoor Environmental Quality for Public Health: Impediments and Policy Recommendations, 115(6) (2007) 953-957.
[5] G. de Gennaro, P.R. Dambruoso, A. Di Gilio, V. Di Palma, A. Marzocca, M. Tutino, Discontinuous and Continuous Indoor Air Quality Monitoring in Homes with Fireplaces or Wood Stoves as Heating System, Int J Environ Res Public Health 13(1) (2015) 78.
[6] S.C. Doll, E.L. Davison, B.R. Painting, Weatherization impacts and baseline indoor environmental quality in low income single-family homes, Building and Environment 107 (2016) 181-190.
[7] J. Langevin, P.L. Gurian, J. Wen, Tracking the human-building interaction: A longitudinal field study of occupant behavior in air-conditioned offices, Journal of Environmental Psychology 42 (2015) 94-115.
[8] T. Deng, X. Shen, X. Cheng, J. Liu, Investigation of window-opening behaviour and indoor air quality in dwellings situated in the temperate zone in China, Indoor and Built Environment (2020).
[9] Y. Geng, B. Lin, J. Yu, H. Zhou, W. Ji, H. Chen, Z. Zhang, Y. Zhu, Indoor environmental quality of green office buildings in China: Large-scale and long-term measurement, Building and Environment 150 (2019) 266-280.
[10] Y. Men, J. Li, X. Liu, Y. Li, K. Jiang, Z. Luo, R. Xiong, H. Cheng, S. Tao, G. Shen, Contributions of internal emissions to peaks and incremental indoor PM2.5 in rural coal use households, Environ Pollut 288 (2021) 117753.
[11] J. Palmisani, A. Di Gilio, M. Viana, G. de Gennaro, A. Ferro, Indoor air quality evaluation in oncology units at two European hospitals: Low-cost sensors for TVOCs, PM2.5 and CO2 real-time monitoring, Building and Environment 205 (2021).
[12] X. Dai, J. Liu, X. Li, L. Zhao, Long-term monitoring of indoor CO2 and PM2.5 in Chinese homes: Concentrations and their relationships with outdoor environments, Building and Environment 144 (2018) 238-247.

Journal Pre-proof

[13] K. Huang, J. Song, G. Feng, Q. Chang, B. Jiang, J. Wang, W. Sun, H. Li, J. Wang, X. Fang, Indoor air quality analysis of residential buildings in northeast China based on field measurements and longtime monitoring, Building and Environment 144 (2018) 171-183.
[14] A. Tunyagi, T. Dicu, K. Szacsvai, B. Papp, G. Dobrei, C. Sainz, A. Cucoş, Automatic system for continuous monitoring of indoor air quality and remote data transmission under SMART_RAD_EN Project, Studia Universitatis Babeș-Bolyai Ambientum 62(2) (2017) 71-80.
[15] P.S. Hui, L.T. Wong, K.W. Mui, Sampling strategies of indoor air quality assessment for offices, Facilities 25(5/6) (2007) 179-184.
[16] A.K.J.A.T. Persily, Evaluating building IAQ and ventilation with indoor carbon dioxide, 103(Pt 2) (1996).
[17] K.W. Mui, L.T. Wong, P.S. Hui, A New Sampling Approach for Assessing Indoor Air Quality, Indoor and Built Environment 15(2) (2016) 165-172.
[18] C.M. Long, H.H. Suh, P. Koutrakis, Characterization of indoor particle sources using continuous mass and size monitors, J Air Waste Manag Assoc 50(7) (2000) 1236-50.
[19] E. Asadi, M.C. da Silva, J.J. Costa, A systematic indoor air quality audit approach for public buildings, Environ Monit Assess 185(1) (2013) 865-75.
[20] China's Ministry of Health, GB/T18883. Indoor Air Quality Standard, 2002.
[21] World Health Organization(WHO), WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005, 2020.
[22] United States Environmental Protection Agency(EPA), Typical Indoor Air Pollutants. , 2009.
[23] M. Mannan, S.G. Al-Ghamdi, Indoor Air Quality in Buildings: A Comprehensive Review on the Factors Influencing Air Pollution in Residential and Commercial Structure, Int J Environ Res Public Health 18(6) (2021).
[24] Indoor Air Quality Management Group, A Guide on Indoor Air Quality Certification Scheme for Offices and Public Places., 2019.
[25] World Health Organization(WHO), Air Quality Guidelines for Europe, 2000.
[26] Institute of Environmental Epidemiology; Ministry of the Environment: Singapore, Guidelines for Good Indoor Air Quality in Office Premises, 1996, pp. 1-47.
[27] National Institute for Occupational Safety and Health(NIOSH), NIOSH Pocket Guide to Chemical Hazards (NPG). 2004.
[28] Health Canada, Residential Indoor Air Quality Guideline, 2007.
[29] Health and Safety Executive, EH40/2005 Workplace Exposure Limits, 2011.
[30] The National Health and Medical Research Council, Goals for Maximum Permissible Levels of Pollutants in Indoor Air. In Interim National Indoor Air Quality Goals; The National Health and Medical Research Council: Melbourne, Australia, 1996.
[31] China's Ministry of Construction, GB-50325-2020. Standard for indoor environmental pollution control of civil building engineering, 2020.
[32] P.K. Cheung, C.Y. Jim, Impacts of air conditioning on air quality in tiny homes in Hong Kong, Sci Total Environ 684 (2019) 434-444.
[33] S. Mentese, N.A. Mirici, T. Elbir, E. Palaz, D.T. Mumcuoğlu, O. Cotuker, C. Bakar, S. Oymak, M.T. Otkun, A long-term multi-parametric monitoring study: Indoor air quality (IAQ) and the sources of the pollutants, prevalence of sick building syndrome (SBS) symptoms, and respiratory health indicators, Atmospheric Pollution Research 11(12) (2020) 2270-2281.
[34] A.Y. Lim, M. Yoon, E.H. Kim, H.A. Kim, M.J. Lee, H.K. Cheong, Effects of mechanical ventilation

Journal Pre-proof

Journal Pre-proof

on indoor air quality and occupant health status in energy-efficient homes: A longitudinal field study, Sci Total Environ 785 (2021) 147324.

[35] S. Abraham, X. Li, A Cost-effective Wireless Sensor Network System for Indoor Air Quality Monitoring Applications, Procedia Computer Science 34 (2014) 165-171.

[36] China's Ministry of Housing and Urban-Rural Development, thermal Design Code for Civil Buildings.(GB 50176-93), China building industry press, 2016.

[37] J. Liu, X. Dai, X. Li, S. Jia, J. Pei, Y. Sun, D. Lai, X. Shen, H. Sun, H. Yin, K. Huang, H. Tan, Y. Gao, Y. Jian, Indoor air quality and occupants' ventilation habits in China: Seasonal measurement and long-term monitoring, Building and Environment 142 (2018) 119-129.

[38] Y. Li, Y.-h. Chiu, L.C. Lu, Energy and AQI performance of 31 cities in China, Energy Policy 122 (2018) 194-202.

[39] Y. Zhou, H. Xiao, H. Guan, N. Zheng, Z. Zhang, J. Tian, L. Qu, J. Zhao, H. Xiao, Chemical composition and seasonal variations of PM2.5 in an urban environment in Kunming, SW China: Importance of prevailing westerlies in cold season, Atmospheric Environment 237 (2020).

[40] Y. Wang, D. Mauree, Q. Sun, H. Lin, J.L. Scartezzini, R. Wennersten, A review of approaches to low-carbon transition of high-rise residential buildings in China, Renewable and Sustainable Energy Reviews 131 (2020).

[41] J. Pei, Y. Yin, J. Liu, Long-term indoor gas pollutant monitor of new dormitories with natural ventilation, Energy and Buildings 129 (2016) 514-523.

[42] C.J. Kahler, B. Sammler, J. Kompenhans, Generation and control of tracer particles for optical flow investigations in air, EXPERIMENTS IN FLUIDS 33(6) (2002) 736-742.

[43] L. Lei, W. Chen, Y. Xue, W. Liu, A comprehensive evaluation method for indoor air quality of buildings based on rough sets and a wavelet neural network, Building and Environment 162 (2019).

[44] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowledge-Based Systems 98 (2016) 1-29.

[45] H. Zhou, Z. Deng, Y. Xia, M. Fu, A new sampling method in particle filter based on Pearson correlation coefficient, Neurocomputing 216 (2016) 208-215.

[46] W.A.N. Joseph Lee Rodgers, Thirteen Ways to Look at the Correlation Coefficient., The American Statistician 42 (1988) 59-66.

[47] A. Houghton, C. Castillo-Salgado, Analysis of correlations between neighborhood-level vulnerability to climate change and protective green building design strategies: A spatial and ecological analysis, Building and Environment 168 (2020).

[48] J.P. Gao, L.P. Ma, Statistics, Capital Economy and Trade University Publishing Company, Beijing (2004) 160-190.

[49] S.C. Sekhar, S.E. Goh, Thermal comfort and IAQ characteristics of naturally/mechanically ventilated and air-conditioned bedrooms in a hot and humid climate, Building and Environment 46(10) (2011) 1905-1916.

[50] M.M.M. Abdel-Salam, Outdoor and indoor factors influencing particulate matter and carbon dioxide levels in naturally ventilated urban homes, Journal of the Air & Waste Management Association 71(1) (2021) 60-69.

[51] N. Jacek, G. Taseusz, K.Z. Bozena, L. Jerzy, Indoor Air Quality (IAQ), Pollutants, Their Sources and Concentration Levels, Building and Environment Vol. 27, No. 3 (1992) No. 3, pp. 339-356.

[52] Y. Zhao, A. Li, R. Gao, P. Tao, J. Shen, Measurement of temperature, relative humidity and concentrations of CO, CO2 and TVOC during cooking typical Chinese dishes, Energy and Buildings

Journal Pre-proof

69 (2014) 544-561.

[53] C. Huang, W. Liu, J. Cai, X. Wang, Z. Zou, C. Sun, Household formaldehyde exposure and its associations with dwelling characteristics, lifestyle behaviours, and childhood health outcomes in Shanghai, China, Building and Environment 125 (2017) 143-152.

[54] A. Stamatelopoulou, D.N. Asimakopoulos, T. Maggos, Effects of PM, TVOCs and comfort parameters on indoor air quality of residences with young children, Building and Environment 150 (2019) 233-244.

[55] Z. Li, Q. Wen, R. Zhang, Sources, health effects and control strategies of indoor fine particulate matter (PM2.5): A review, Sci Total Environ 586 (2017) 610-622.

[56] Y. Zhao, H. Sun, D. Tu, Effect of mechanical ventilation and natural ventilation on indoor climates in Urumqi residential buildings, Building and Environment 144 (2018) 108-118.

[57] M. Guo, X. Pei, F. Mo, J. Liu, X. Shen, Formaldehyde concentration and its influencing factors in residential homes after decoration at Hangzhou, China, Journal of Environmental Sciences 25(5) (2013) 908-915.

[58] T. Godish, J. Rouch, Mitigation of residential formaldehyde contamination by indoor climate control, American Industrial Hygiene Association journal 47 (1986) 792-797.

[59] C.R. Frihart, J.M. Wescott, T.L. Chaffee, K.M. Gonner, Formaldehyde Emissions from Urea-Formaldehyde- and No-Added-Formaldehyde-Bonded Particleboard as Influenced by Temperature and Relative Humidity, Forest Products Journal 62 (2012) 551-558.

[60] L. Zhao, J. Liu, J. Ren, Impact of various ventilation modes on IAQ and energy consumption in Chinese dwellings: First long-term monitoring study in Tianjin, China, Building and Environment 143 (2018) 99-106.

Journal Pre-proof

Journal Pre-proof

Highlights

- A systematic approach was presented for IEQ data sampling.
- Sampling strategy, frequency and required number of dwellings were studied.
- Discrete sampling strategy achieves better performance.
- An algorithm was proposed to calculate sampling frequency of IEQ parameters.
- Required number of dwellings depends on the variable coefficient of parameters.

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof