Zhao-Ge Liu, Xiang-Yang Li, Grunde Jomaas,

Effects of governmental data governance on urban fire risk: A city-wide analysis in China,

International Journal of Disaster Risk Reduction,

Volume 78,

2022,

103138,

ISSN 2212-4209

The article has been published in final form at:

https://doi.org/10.1016/j.ijdrr.2022.103138

Effects of Governmental Data Governance on Urban Fire Risk: A City-Wide Analysis in China

Abstract

The effects of data governance (as a means to maximize big data value creation in fire risk management) performance on fire risk was analyzed based on multi-source statistical data of 105 cities in China from 2016 to 2018. Specifically, data governance was first quantified with ten detailed indicators, which were then selected for explaining urban fire risk through correlation analysis. Next, the sample cities were clustered in terms of major socio-economic characteristics, and then the effects of data governance were examined by constructing multivariate regression models for each city cluster with ordinary least squares (OLS). The results showed that the constructed regression models produced good interpretation of fire risk in different types of cities, with coefficient of determination (R^2) in each model exceeding 0.65. Among the indicators, the development of infrastructures (e.g. data collection devices and data analysis platforms), the level of data use, and the updating of fire risk related data were proved to produce significant effects on the reduction of fire frequency and fire consequence. Moreover, the organizational maturity of data governance was proved to be helpful in reducing fire frequency. For the cities with large population, the cross-department sharing of high-value data was found to be another important determinant of urban fire frequency. In comparison with existing statistical models which interpreted fire risk with general social factors (with the highest R^2 = 0.60), these new regression models presented a better statistical performance (with the average R^2 = 0.72). These findings are expected to provide decision support for the local governments of China and other jurisdictions to facilitate big data projects in improving fire risk management.

Keywords: Urban fire risk; fire risk management; big data technologies; data governance; socio-economic factors; city-wide analysis

1 Introduction

Fires in urban settings are often accompanied by serious consequences. There are many different sources for the consequences of fires, and there can be quite a spread in reported numbers for fires, fatalities and monetary losses associated with fires. According to the statistics published by the International Association of Fire and Rescue Services (CTIF), more than 8 million fire incidents occur per year worldwide, resulting in more than 120,000 deaths [1, 2]. In US alone, fires caused 3,704 deaths and $14.8 billion direct property losses in 2019 [3]. The direct property loss refers to the direct loss of house, structure, equipment, and other property. The causes of the direct loss include burning, smoking, radiation, and demolition, collision, water stains, and pollution during the firefighting [4]. China is also one of the countries that are heavily affected by fire disasters. According to the data from the Fire and Rescue Department of China, a total of 252,000 fire incidents occurred in mainland China in 2020, resulting in 1,183 deaths and $621 million direct property losses [5].

The threat of fire is generally treated in three ways. The first is fire safety design considering the fire dynamics under physical conditions (e.g., materials, structures, and fire protection measures) before building construction [6, 7]. The second is fire risk management (e.g., fire hazard identification, fire risk prediction, and the maintenance of fire equipment) after buildings have been in place [8, 9]. The third one is the rescue and evacuation procedures in fire emergencies, which need to consider both fire dynamics and human behavior [10]. Due to the high density of buildings in use and large population in cities, a growing concern for urban fire departments is how to reduce the frequencies and consequences of urban fires through effective fire risk management [11]. In fire risk management, fire risk analysis plays a fundamental role as it provides evidence of the necessity of reducing fire risk and the choice of fire risk treatment measures [12].

Traditional fire risk analysis has generally interpreted fire risk from a physical perspective. Specifically, researchers tried to identify the potential associations between fire risk level and various physical characteristics (e.g., building characteristics, state of implemented fire safety measures, surroundings, weather conditions, and topographical features). Further, the suggestions for enhancing fire risk management were provided according to those obtained associations [13, 14]. With the rapid development of cities, it has been gradually found that urban fire risk is not only determined by the physical environment but is also strongly associated with human behaviors (e.g. building occupants). As a result, more and more studies have been proposed that try to interpret urban fire risk from a social perspective. For example, a large population size generally indicates high fire risk due to more risk behaviors (e.g. playing with fire, unsafe cooking behaviors, and occupying fire exits) that can lead to or aggravate fire incidents [4]. Other socio-economic and demographic determinants (e.g., population density, income level, per capita GRP, industrial development, and household type) have also been proposed to influence the fire risk [15-17].

However, most of the existing studies of fire risk analysis have focused on the fire behavior and the behavior of building occupants, and have to a lesser extent considered the intervention behaviors of urban fire departments in reducing fire risk. This gap greatly limits the discovery and adoption of effective measures for fire risk treatment [18, 19]. Under such a situation, it is important to continually discover emerging endeavors of urban fire departments and analyze their effects on improving fire risk management and reducing the fire risk. In particular, with the development of big data technologies, e.g. artificial intelligence (AI), visual reality (VR), and Internet of Things (IoT), more and more cities have explored various big data projects to improve fire risk management by utilizing big data. In Atlanta, USA, for example, various data (e.g., structures, materials, and population density) was used with a machine learning (ML) algorithm to find buildings with high fire risk [20]. In Suzhou, China, the city fire department built a fire prediction system called 'Fire Eye' in 2015, which can dynamically predict the fire incidence of each building using multi-source data including building structure, population size, electricity, and water consumption [21].

Although big data has obvious benefits on fire risk management, big data projects have been proven successful in only a small number of cities. In particular, the use of big data faces various data-related problems. For example, big data analysis often requires data from different governmental sectors, but the data sharing is prone to be limited partly due to organizational coordination difficulties and the lack of unified big data management platforms [22, 23]. Moreover, the collection and processing of some types of big data (e.g., video data and textual data) require specific digital infrastructures (e.g. data centers, cloud platforms, and video monitoring devices) [24, 25]. Typically, it would be too precarious to develop big data projects with reliable performance without considering data-related matters, e.g., data collection, data sharing, data processing, and data use. Under such a situation, urban fire departments are turning to data governance as a means to maximize data value creation in fire risk management, which involves the effective collection, management, and use of big data [26, 27]. Generally, data governance refers to a series of governance activities that enable organizations to ensure that high quality exists throughout the complete big data lifecycle, and data controls are implemented to support business objectives [28]. In the fire risk management context, data governance mostly works through high-quality data supply and effective data management for improving fire services [29]. For example, the Guiyang city fire brigade is using IoT to remotely monitor possible fire risks with the various data (e.g., temperature, electricity current, and voltage) provided by the near field communication (NFC) chips that are installed in hundreds of facilities, such as public hospitals and conference centers [30].

Previous studies of data governance tend to show the potential value of specific ways to use data (e.g., computer vision and GIS), but overlook a wide range of other data governance work (e.g., data sharing strategies, data quality management, and infrastructure construction) [31]. A few studies have investigated the effects of data governance on fire risk in a relatively comprehensive way. However, their effects in these studies were often only qualitative and were generally focused on a specific city, and as such, had limited generality [32, 33].

To show how data governance affects fire risk, China was chosen as an example due to the large-scale actions of data governance and their intensive uses in fire risk management. In 2017, the national fire department of China published a document that required prefecture-level and above cities to realize preliminary data governance of fire risk management before 2018. Since then, Chinese city governments have successively established their own data management platforms for fire risk management, which enables the unified and standardized governance of data [29]. Based on the multi-source data provided by the platforms, successful cases of big data projects have been emerging, such as the 'Fire Eye' system in Suzhou and the 'Smart Fire' system in Guiyang [34]. The intensive development of data governance helps highlight its effects.

The effects of data governance on disaster risk are often assessed in the literature using a qualitative approach or it is simply documented [29]. The reason most often raised is the lack of relevant data that could help to conduct a more rigorous analysis [35]. To obtain the general conclusions of interpreting disaster risk, a multivariate regression method is most common due to its advantages in rigorous and quantitative model construction and validation [35, 36]. The choice of the model depends on the type of the dependent variable used to capture the outcome. In this paper, the regression model with ordinary least squares (OLS) was adopted because the dependent variables (describing the fire risk) were all continuous ones [37].

The biggest challenge when studying the effects of data governance is that the analysis of some indicators is rather difficult due to the limited availability of city-level indicator data (e.g. data collection ways, infrastructure construction, and data use details). To achieve a comprehensive analysis of data governance conditions in each city, multi-source materials (e.g., *China Governmental Data Governance Reports* [38], *China Data Opening Reports* [39-41], governmental documents, and statistical yearbooks of each sample city), and city-level open data platforms (see the Appendix for details) were surveyed, which supports the statistical analysis of more than 100 Chinese cities.

At the same time, many data governance indicators (e.g., data governance organization and policy support) are qualitatively described in previous studies [42] and are challenging to use directly for building statistical models. In fact, most of these indicators can be measured quantitatively by dividing the values into different levels. For example, organization structure can be measured with a maturity degree using a capability maturity model [43, 44]. In the current study, a quantitative indicator framework is provided to support statistical analysis.

Another challenge is that some other socio-economic factors (e.g., population and GDP) have strong effects on urban fire risk in China [4, 15-17], which may cause a bias of statistical results and thus affect the conclusions. To figure out the effects of data governance, the best way is to incorporate these influential socio-economic factors as control variables in constructed statistical models. Specifically, sample cities are grouped in terms of the control variables through clustering analysis [4]. Further, the control variables are incorporated into the final models to adjust the effects of data governance and obtain conclusions regarding different types of cities. Compared to the commonly used K-means clustering method, the hierarchical clustering analysis adopted in this paper does not require to pre-specify the number of clusters [45]. This is helpful for finding the optimal clustering results.

It should be noted that data governance statistics may be more beneficial for the fire risk management of cities with high urbanization and good fire prevention due to the limitations of current big data technologies. On the one hand, current big data technologies generally have weak performance in dealing with risks brought by indoor human behaviors, as shown in Table 1. Therefore, for cities with poor fire prevention, data governance is likely to contribute less to fire risk management because the main concern in these cities is generally the reduction of indoor fire risk. On the other hand, current big data technologies rely on various infrastructures and devices for collecting required data. However, in the rural areas of cities where there are fewer buildings and less population, it is both unnecessary and unrealistic to have infrastructures and devices in every part of the areas. Hence, for those areas, data governance might be less effective due to the lack of relevant data.

Table 1 The advantages and disadvantages of current big data technologies in different fire risk management tasks. The tasks were derived from existing research [11] and a practical survey of more than 30 Chinese cities.

Fire risk management tasks	Advantages of big data technologies	Disadvantages of big data technologies
Fire risk identification	• Real-time identification with sensors • Identifying various fire hazards • Precise identification results	• Difficult to identify risks that are related to indoor human behaviors • Relying on infrastructures and devices
Fire risk analysis	• Comprehensive analysis with integration of multi-source data • Discovering useful fire risk indicators based on real world data	• Data management difficulties including data collection, sharing and security • Difficult to analysis risks about indoor human behaviors due to the lack of relevant data
Fire risk evaluation	• Efficient evaluation and reliable results • Dynamic evaluation with real-time data sharing	• Some indicators cannot be quantified and thus cannot be used for evaluation • Precision depends on data amount
Fire risk treatment	• Interactive and real-time communication • Joint treatment by different departments	• Relying on infrastructures and devices • Data management difficulties

This study contributes to the existing literature in two ways. First, this study extends the urban fire risk research by delineating data governance activities as fire departments' performance indicators shaping the reduction of fire risk across different cities. Previous research on urban fire risk has primarily focused on the behaviors of building occupants from the socio-economic perspective [4, 16], which lacks analysis of the powerful endeavors of urban fire departments. The findings of this study indicate that data governance facilitates effective fire risk management and enriches the interpretation of urban fire risk. Second, this study contributes to the data governance research by extracting and quantifying the governmental data governance indicators under the fire risk management context from a comprehensive perspective. Previous studies tend to either interpret data governance from specific aspects (e.g., data use) or analyze the comprehensive data governance performance in a specific region [31, 32]. In this study, data governance has been analyzed from various aspects (e.g., infrastructures, policies, and organizations) based on a large-scale and cross-city investigation in China.

In summary, it was investigated whether and how governmental data governance affects urban fire risk based on statistical data of 105 Chinese cities from 2016 to 2018 (to obtain general conclusions). In this process, regression models with OLS were employed to identify the data governance determinants of fire risk. To support the statistical analysis, data governance indicators were extracted from a wide range of data sources and then expressed in a quantitative way for establishing a quantitative indicator framework. Meanwhile, a hierarchical clustering method was used to group cities in terms of socio-economic indicators for investigating the applicability of data governance in different types of cities. Finally, the major findings and their applicability, implications, and limitations are discussed.

2 Data and methods

A statistical method was used to investigate the effects of data governance on fire risk. The first step of the method was to build a comprehensive indicator framework that integrated both data governance and major socio-economic features. Next, data governance indicators that have high correlation degrees with fire risk were selected as explanatory variables, while the socio-economic indicators were selected as control variables. The cities were then grouped in terms of socio-economic indicators using hierarchical clustering. Finally, multivariate regression models with OLS for different types of cities were employed to identify the data governance determinants of fire risk. Here, fire risk was evaluated with two indicators, i.e. fire frequency and fire loss [4]. Among them, fire frequency was the total number of fire incidents in a year and fire loss was assessed by annual direct property loss (which could be obtained in the China Fire Yearbook) [46-48]. The flow chart of the methodology is shown in Fig. 1. It should be noted that the three city clusters in the figure are just examples and the final number of the city clusters depends on the statistical analysis.

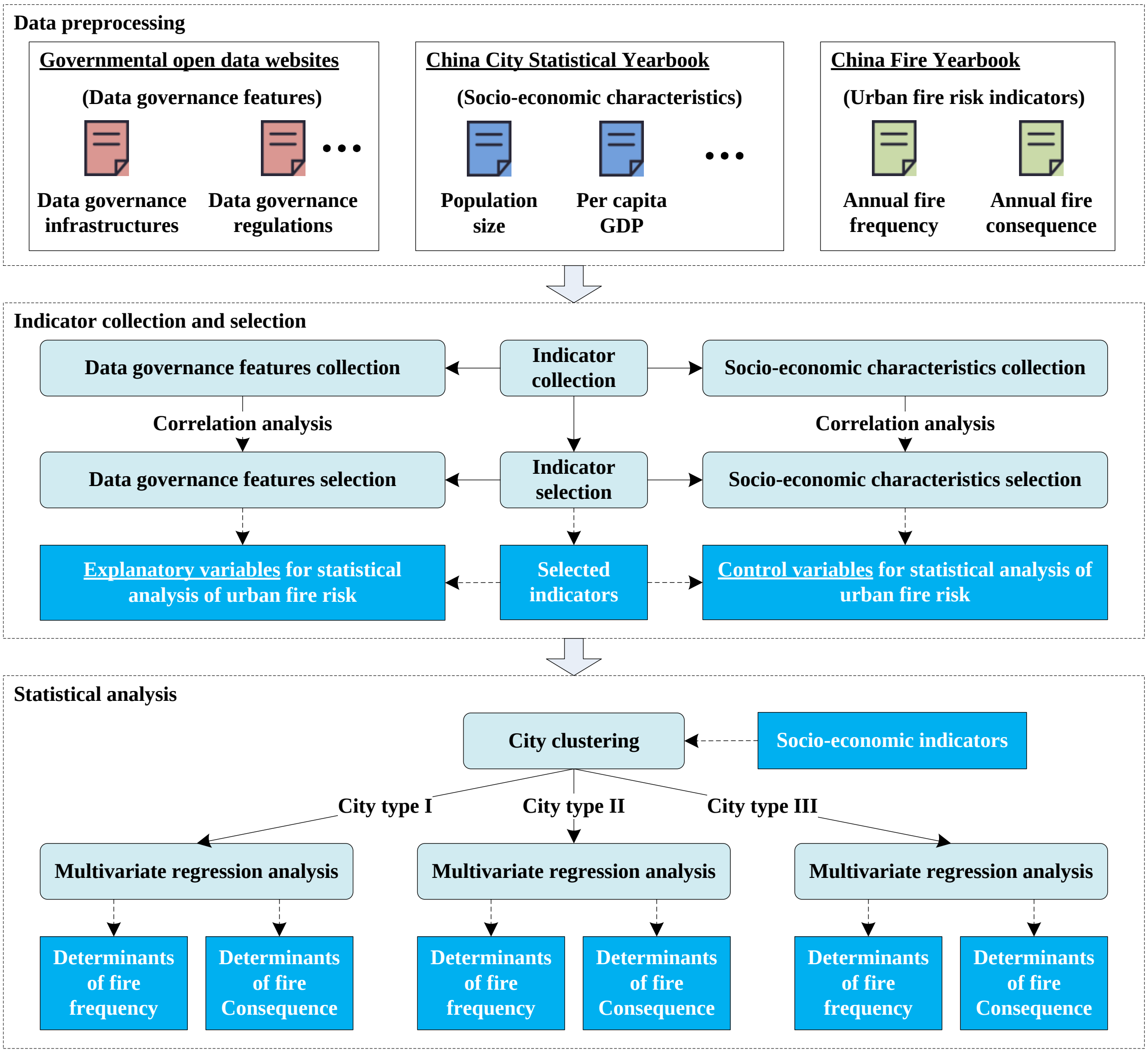

Fig. 1 Flow chart of the methodology.

2.1 Data sources

This study investigated 105 Chinese cities of different administrative levels, including 21 provincial capitals, 74 prefecture-level cities, and 10 county-level cities. Among them, provincial capital is the center of a province (i.e., first-level administrative region) in various aspects such as politics and economy. Prefecture-level city is the second-level administrative region, ranking below a province and above a county in China. County-level city is the third-level administrative region in China and is generally governed by a prefecture-level city. Other cities in China were not selected for analysis, mainly due to their non-significant development of data governance and to the incompleteness of data governance statistics for the following analysis. The spatial distribution of the selected cities is shown in Fig. 2. As the sample covered most of mainland China, it can be deemed representative for analyzing the effects of data governance.

Three main types of data were included in this study.

(1) The indicator values of data governance conditions were derived from governmental open data platforms (which are listed in the Appendix) and other related sources, e.g. *China Governmental Data Governance Reports* and *China Data Opening Reports* [38-41]. Among them, the governmental open data platforms provide the details of more than 160 cities regarding the infrastructure development, data

sharing, data updating, policy support, and organizational maturity of fire-related data governance. The *China Governmental Data Governance Reports* have included the data security conditions as well as some abovementioned indicators of more than 200 Chinese cities. The *China Data Opening Reports* focus on data use and data standardization, and have published relevant data of nearly 130 cities. It should be noted that the publishers of the data governance conditions were either governments or major research organizations in China, and thus the quality of the collected data was reliable.

(2) Data related to fire risk (i.e. fire frequency and fire loss) was derived from *China Fire Yearbook* from 2016 to 2018 [46-48]. The *China Fire Yearbook* (published by the Fire Service Bureau) reports the fire statistics of more than 270 cities in China, which ensures the completeness and accuracy of the data. Since fire statistics can vary from year to year, the outliers (e.g., the high numbers of fire incidents linked to other disasters such as draught) in the data were detected and removed by analyzing the Z-score and the annual reports of the fire departments [49].

(3) Control variables: major socio-economic variables proposed in previous studies. These data came from the *China City Statistical Yearbook* from 2016 to 2018 [50-52]. The *China City Statistical Yearbook* is an official document that shows the complete socio-economic conditions (e.g., population size and per capita GDP) of nearly all Chinese cities. Previous fire risk research has used the dataset to analyze the effects of major socio-economic factors on urban fire risk levels [4].

During the data collection, the missing data accounted for less than 3% (3 cities) of the total (105 cities), and was addressed with the average interpolation [53]. Before performing statistical analysis, the abovementioned data was normalized to avoid the effects of unit conflicts. Data scaling was adopted to do this by converting the indicators into no-dimension variables ranging from -1 to 1. Using a mean normalization method [4], the equation for data scaling can be illustrated as follows.

$$x^* = \frac{x - \overline{x}}{x_{\max} - x_{\min}} \tag{1}$$

where x^* represents the indicator value after data scaling, x is the original indicator value, $\overline{x}$ is the mean of indicator values, $x_{\max}$ denotes the maximum of indicator values and $x_{\min}$ denotes the minimum of indicator values.

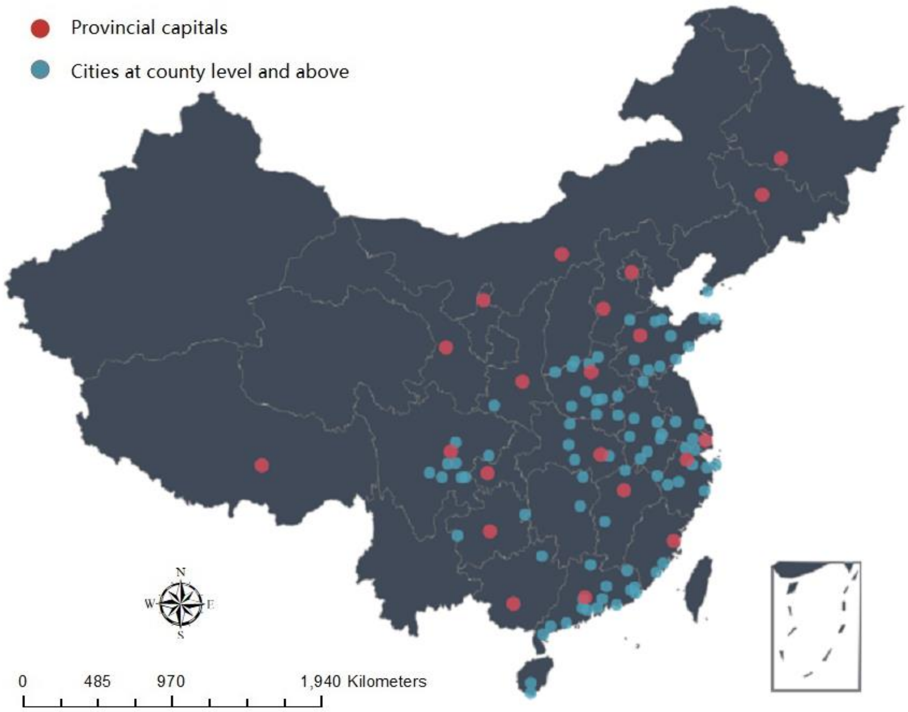

Fig. 2. The spatial distribution of the selected cities

2.2 Indicator collection and selection

To obtain the data governance determinants of fire risk based on the previous research on data governance and intelligent fire risk management [42, 44, 54-57], a relatively comprehensive indicator framework that considered five wide-ranging indicators was established. These indicators were; 1) infrastructure development, 2) organizational maturity, 3) policy support, 4) data sharing (and related data updating, standard, and security), and 5) data use. The indicators were then measured with 10 statistical items, as shown in Table 2. It should be noted that to highlight the data governance effects on urban fire risk, the indicators mainly refer to the development of fire risk management instead of the general conditions. For example, the organizational maturity describes the organizational structure of the data governance regarding fire risk management. Further descriptions of the indicator groups are given in the following.

Table 2 Indicators used in this study and corresponding explanations

Indicator type	Indicator	Reference	Statistical items	Index
Data governance features (Explanative variables)	Infrastructure development	[42]	Expenditure on built infrastructures	X_1
	Organizational maturity	[42, 44, 54]	Level of organizational structure	X_2
	Policy support	[42, 44]	Number of policies about data governance	X_3
	Data sharing	[54]	Number of shared datasets	X_4
	High-value data sharing	[54]	Number of shared high-value datasets [a]	X_5
	Data updating	[55]	Number of updated data [b]	X_6
	Data standard	[44, 54, 55]	Number of meta-data standards	X_7
	Data use	[44, 54, 55]	Number of data sources for decision support	X_8
		[44, 54, 55]	Level of data use	X_9
	Data security	[54, 56]	Level of data security	X_{10}
Socio-economic characteristics (Control variables)	Population size	[4, 16]	Registered population at year end	X_{11}
	Per capita GDP	[4, 16]	Per capita Gross Regional Product	X_{12}
	Industrial development level	[4, 16]	Number of industrial enterprises	X_{13}
	Income level	[4, 16]	Average wage of employed staff and workers	X_{14}
Urban fire risk	Fire frequency	[4]	Annual number of urban fires	Y_1
	Fire consequence	[4]	Annual amount of fire direct property losses	Y_2

[a] High-value data: the data with annual downloads exceeding 10,000.
[b] Updated data: the data with average updating frequency exceeding once every three months.

(1) **Infrastructure development.** Intelligent fire risk management and other e-government services rely on the various types of data that is collected using devices (e.g. video monitors and smoke detectors) and analyzed with the support of infrastructures (e.g. data use platforms and data storage centers) [42]. The expenditure on these data governance infrastructures and devices was used to quantify the infrastructure development level.

(2) **Organizational maturity.** Data governance organizations work for the coordination of different departments to promote data sharing across the departments through data services such as establishing open data platforms and making data regulations [54]. These services help fire departments obtain multi-source data, which is of vital importance for realizing more precise fire risk analysis. Here, the organizational factor was measured using the organizational maturity framework [43, 44], which divides organizational maturity into four levels: Level 1 – Fire departments are responsible for data governance; Level 2 – There are exclusive data governance organizations but they are subordinate to other governmental departments (e.g. information technology department); Level 3 – There are exclusive data governance organizations which have the same administrative level with other departments; Level 4 – Based on level 3, data governance organizations collaborate with external organizations (e.g. companies and nongovernmental organization) to facilitate data

aggregation. For example, Xiamen city has an exclusive data governance organization (i.e., Xiamen Big Data Bureau) but it is subordinate to the Industry and Information Technology Bureau. Hence, the organization maturity of Xiamen belongs to Level 3.

(3) **Policy support.** Data governance policies provide high-level guidelines and rules regarding the creation, acquisition, storage, security, quality, and permissible use of data [42, 44]. With the support of these policies, it is more efficient for fire departments as well as other collaborative organizations to communicate key objectives and conduct data governance activities for intelligent fire risk management.

(4) **Data sharing and related data updating, standard and security.** The number of available multi-source datasets is the most direct indicator for evaluating data sharing performance [54]. In addition, data updating influences the timeliness and precision of the shared data [57]. Data standard helps reduce the ambiguity and structural conflicts of data from different sources [54]. Data security techniques protect the shared data from destructive forces and from unwanted actions, e.g. a cyberattack or a data breach [42].

(5) **Data use.** The effects of data governance depend heavily on how the relevant data is used [54, 55]. In urban fire risk management, the level of data use can be divided into four levels: Level 1 – There is no exclusive data use platform for fire risk management; Level 2 – There are only platforms for specific tasks in fire risk management, e.g. fire hazard identification and fire risk treatment; Level 3 – There are platforms working for the multiple tasks in fire risk management but only use the data from fire departments; Level 4 – Based on Level 3, there are unified platforms which integrate data from other organizations. For example, Suzhou fire department has built a unified platform (i.e. the 'Fire Eye') for various fire services that involves features such as hazard identification and fire risk assessment based on multi-source data (e.g., electricity, construction, and population). As the Suzhou government has a unified platform working for the multiple tasks and integrates multi-source data, its data use belongs to Level 4.

In addition to the data governance features, major socio-economic characteristics were identified from previous studies [4, 16]. These indicators were employed as control variables as some of them produce strong effects on fire risk.

After the indicator data collection, a correlation analysis between each indicator and fire risk was carried out to select the highly correlated indicators for establishing statistical models. The correlation degree was measured with the Person correlation coefficient [58], with the equation being as follows.

$$\rho\left(X_m, Y_n\right) = \frac{Cov\left(X_m, Y_n\right)}{\sqrt{Var\left[X_m\right]Var\left[Y_n\right]}} \tag{2}$$

Here X_m denotes the explanative and control variable m (i.e. one of the abovementioned variables), Y_n denotes the dependent variable n (i.e. fire frequency and fire consequence), Cov (X_m, Y_n) is the covariance of X_m and Y_n, Var [X_m] and Var [Y_n] represent the variance of X_m and Y_n, respectively. The correlation degree ρ ranges from -1 to 1 where -1 denotes absolute negative correlation and 1 denotes absolute positive correlation. In addition, the results of correlation analysis were statistically tested using p-value. Generally, the p-value of the selected variables should not exceed 0.05.

2.3 City clustering

In city clustering, the cities were grouped into different clusters to investigate the effects of data governance in different types of cities. Specifically, the cities with similar socio-economic characteristic(s) were grouped into one city cluster using a similarity measure, with different clusters of cities being highly dissimilar. By clustering cities, corresponding multivariate regression models can be established to better understand the applicability of data governance statistics.

The set of cities is expressed as **C** = {**C**$_1$, **C**$_2$,…, **C**$_n$}. Each city is expressed as a d-dimension vector, i.e. **C**$_i$ = {m_{i1}, m_{i2},…, m_{id}}, where m_{i1}, m_{i2},…, m_{id} are the socio-economic characteristic(s) for measuring the similarity. The cities are then clustered using hierarchical clustering, which is illustrated with the steps given in the following [45].

Step 1: Initial clustering. Set k = 0 for indicating the number of clusters is $n - k = n$ at the current stage. Hence, each city will be regarded as one cluster, i.e. $\mathbf{G}_i^{(k)} = \{\mathbf{C}_i\}(i=1,2,...,n)$, where $\mathbf{G}_i^{(k)}$ is the set of clusters.

Step 2: Calculating the distance between clusters. Euclidean distance is used to calculate the distance of the cities within same clusters, and average distance is employed to the distance between clusters. The equations for the abovementioned calculation are

$$D(\mathbf{G}_i^{(k)},\mathbf{G}_j^{(k)}) = \frac{1}{n_i n_j}\sum_{\mathbf{C}_p\in\mathbf{G}_i^{(k)}}\sum_{\mathbf{C}_q\in\mathbf{G}_i^{(k)}} d(\mathbf{C}_p,\mathbf{C}_q) \tag{3}$$

$$d(\mathbf{C}_p,\mathbf{C}_q) = \left[\sum_{j=1}^{d}\left(m_{pj}-m_{qj}\right)^2\right]^{\frac{1}{2}} \tag{4}$$

where $D(\mathbf{G}_i^{(k)},\mathbf{G}_j^{(k)})$ is the distance between clusters, $d(\mathbf{C}_p,\mathbf{C}_q)$ is the distance of the cities within clusters, n_i and n_j are the numbers of city clusters $\mathbf{G}_i^{(k)}$ and $\mathbf{G}_i^{(k)}$, respectively. Following this, a symmetric matrix representing the distance among clusters is generated and expressed as $\mathbf{D}^{(k)} = D(\mathbf{G}_i^{(k)},\mathbf{G}_j^{(k)})_{h\times h}$, where h is the number of clusters.

Step 3: Combining city clusters. Searching for the minimal element in $\mathbf{D}^{(k)}$, which is found to be the distance between $\mathbf{G}_i^{(k)}$ and $\mathbf{G}_i^{(k)}$. Then, the two clusters are combined, and new cluster set is generated, which is expressed as $\mathbf{G}_i^{(k+1)}(i=1,2,...,h-1)$. Set $k = k + 1$, $h = h - 1$.

Step 4: Checking the number of clusters. If $h > 2$, the clustering can be continued and the process turns to Step 2. Otherwise, the clustering will be stopped.

Step 5: Evaluating clustering validity. Clustering validity $CV^{(k)}$ is measured using the Calinski-Harabaz Index, i.e. the ratio of inter-class differences $WC^{(k)}$ to intra-class differences $BC^{(k)}$, i.e. $CV^{(k)} = BC^{(k)} / WC^{(k)}$. The equations of the abovementioned indicators are given in Equations 5-7 [45].

$$WC^{(k)} = \sum_{i=1}^{h} WC^{(k)}(\mathbf{G}_i^{(k)}) = \sum_{i=1}^{h}\sum_{\mathbf{C}_j\in\mathbf{G}_i^{(k)}} d(\mathbf{C}_j,\mathbf{r}_i)^2 \tag{5}$$

$$\mathbf{r}_i = \frac{1}{n_i}\bigoplus_{\mathbf{C}_j\in\mathbf{G}_i^{(k)}}\mathbf{C}_j = \frac{1}{n_i}\sum_{\mathbf{C}_j\in\mathbf{G}_i^{(k)}}\mathbf{C}_j \tag{6}$$

$$BC^{(k)} = \sum_{1\le i\le j\le h} d(\mathbf{r}_i,\mathbf{r}_i)^2 \tag{7}$$

Step 6: Generating clusters. For different values of k, evaluating clustering validity using Eqs. (3) – (5). The clustering results with high $CV^{(k)}$ will be selected, and thus the number of clusters is $h = n - k$.

2.4 Establishing multivariate regression models

Multivariate regression models were constructed with ordinary least squares (OLS) for each cluster to provide evidence for interpreting the relationships between data governance determinants and urban fire risk. Specifically, the multivariate regression models are expressed as follows [4].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad \varepsilon \sim N\left(0,\sigma^2\right) \tag{8}$$

where Y denotes the dependent variable (i.e. fire frequency and fire consequence), $X_1, X_2,..., X_k$ are the explanative variables (i.e. the selected data governance indicators), $\beta_1,\beta_2,...,\beta_k$ are corresponding regression coefficients, ε is an error item which obeys normal distribution. In the regression models, the selected socio-economic variables were also incorporated to control their strong effects on fire risk.

The data samples were divided into two datasets that were used for the model training and the model test, respectively. The data from 2016 to 2017 was used as the training data, with the data of 2018 used for the model test. At the stage of model training, regression models were generated using OLS, and the p value of t-test was used to analyze the significance of each regression coefficient. In general, a coefficient is proved significant if its p value is lower than 0.1 [59].

The coefficient of determination (R^2) and root mean squared error (RMSE) were used to test the goodness of fit of the training models [59]. Among them, R^2 ranges from 0 to 1. The closer R^2 is to 1, the higher the fitting degree of regression model is. RMSE indicates the goodness of fit by measuring the differences between the values predicated and those observed. RMSE is non-negative, and the closer RMSE is to 0, the better the goodness of fit.

3 Results

3.1 Data governance feature selection

Table 3 shows the results for the Pearson correlation coefficients of each indicator and fire risk that were calculated with Eq. (1) and with the p-value that was generated with the correlation analysis in SPSS software.

According to the results, infrastructure development (X_1), organizational maturity (X_2), high-value data sharing (X_5), data updating (X_6), data standard (X_7) and data use (X_9) were highly correlated with fire risk and were selected as explanative variables. In addition, population size (X_{11}) was selected as control variable. Policy support (X_3) was not significantly correlated with fire risk partly because current data governance policies focus more on data risk management, but not on data value creation, and thus decision makers have not been motivated to aggregate more data for fire risk management. In the aspect of data sharing, only the sharing of high-value data produced significant effects on fire risk, which explained the importance of selecting this kind of data for effective fire risk management. The results also illustrated that decision makers should pay more attention on the level of data use instead of the number of data sources as the effects of the number of data sources (X_8) on fire risk were not significant. The effects of data security (X_{10}) were also not significant, which can be understood as data security does not directly contribute to data value creation.

Table 3
Results for the correlation analysis based on the data from 2016 to 2018 for 105 cities. Y_1 represents fire frequency, Y_2 represents fire consequence, $X_1 \sim X_{10}$ represent data governance indicators (i.e. explanative variables) and $X_{11} \sim X_{14}$ represent socio-economic indicators (i.e. control variables). The positive numbers indicate positive correlations and the negative numbers indicate negative correlations.

Indicator	Y_1	Y_2	Indicator	Y_1	Y_2
X_1 (Infrastructure development)	-0.52**	-0.51**	X_8 (Data use: Data sources)	-0.23	-0.21
X_2 (Organizational maturity)	-0.69**	-0.57**	X_9 (Data use: Data use level)	-0.47**	-0.62**
X_3 (Policy support)	0.14	0.12	X_{10} (Data security)	-0.27	-0.26
X_4 (Data sharing)	0.08	0.01	X_{11} (Population size)	0.86**	0.78**
X_5 (High-value data sharing)	-0.59**	-0.37**	X_{12} (Per capita GDP)	0.36**	0.30*
X_6 (Data updating)	-0.32*	-0.43**	X_{13} (Industrial development level)	0.30*	0.33*
X_7 (Data standard)	-0.66**	-0.45**	X_{14} (Income level)	0.21	0.26

$^{*}p < 0.1$, $^{**}p < 0.05$

Further, the scatterplots of fire frequency and the selected data governance variables were analyzed to decide the expression functions of the variables that can produce the best goodness of fit, as shown in Figs. 3 and 4. A total of three types of functions were considered for this, namely, linear function, polynomial function and logarithm function. Among the three types of functions, the logarithm function achieved the highest R^2. Hence, these variables were converted into the logarithmic form. As such, for the variable X_i, the converted variable employed in the regression models should be $\ln(X_i)$.

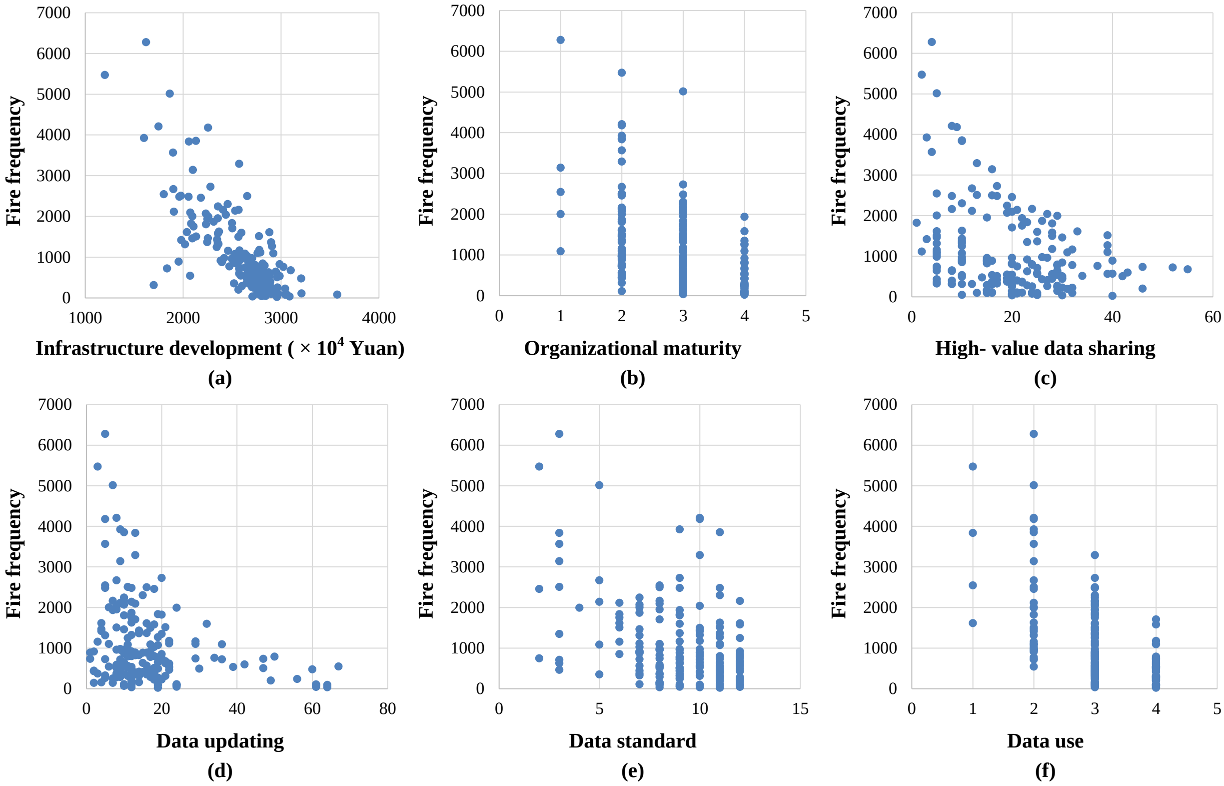

Fig. 3. The scatterplots of fire frequency and data governance features based on the data for 105 cities from 2016 to 2018. Three types of fitting functions are considered for producing the best goodness of fit, i.e. linear function, polynomial function and logarithm function. After comparison, the logarithm function achieved the highest goodness of fit according to the coefficient of determination (R^2).

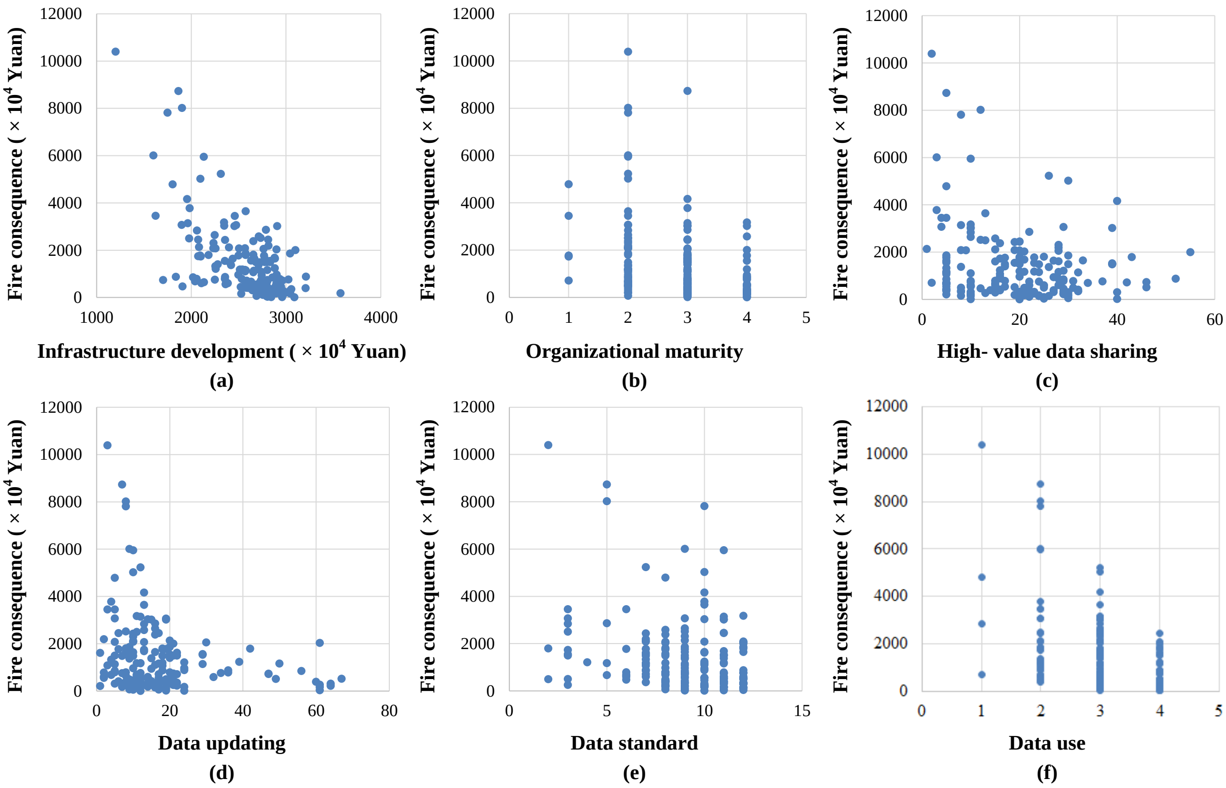

Fig. 4. The scatterplots of fire consequence and data governance features based on the data for 105 cities from 2016 to 2018. Three types of fitting functions are considered for producing the best goodness of fit, i.e. linear function, polynomial function and logarithm function. After comparison, the logarithm function achieved the highest goodness of fit according to the coefficient of determination (R^2).

3.2 Clustering Results

The characteristics for measuring the similarity between cities can be one or more of the four socio-economic indicators shown in Table 1, and there are therefore a number of potential clustering solutions. To select the best solution, the clustering validity under each potential solution was evaluated and compared using Eqs. 3-5. Finally, it was found that choosing the indicator ‘population size’ resulted in the best clustering performance, which was also supported by the correlation analysis results. Consequently, ‘population size’ was used to cluster cities.

Furthermore, the number of clusters was determined through comparing the clustering validity values under each condition, as shown in Fig. 5. It can be seen that the clustering performance under different cluster numbers was pretty similar. However, when there are more than two clusters, the number of cities in each cluster becomes very small, which is prone to cause data overfitting in the following statistical analysis. To avoid this, the cities were finally divided into two clusters, i.e. large cities (with higher population) and medium-sized and small cities (with lower population). Fig. 6 provides the helpful visual information of understanding the features of the two types of cities. Specifically, the threshold between large cities and smaller cities was around 8 million people, and the cluster center cities were Zaozhuang (medium-sized and small cities) and Wuhan (large cities). The distribution of the clustered cities is shown in Fig. 5. Generally, most of the sample cities had low levels of population (with only 30 of the 105 sample cities having higher population).

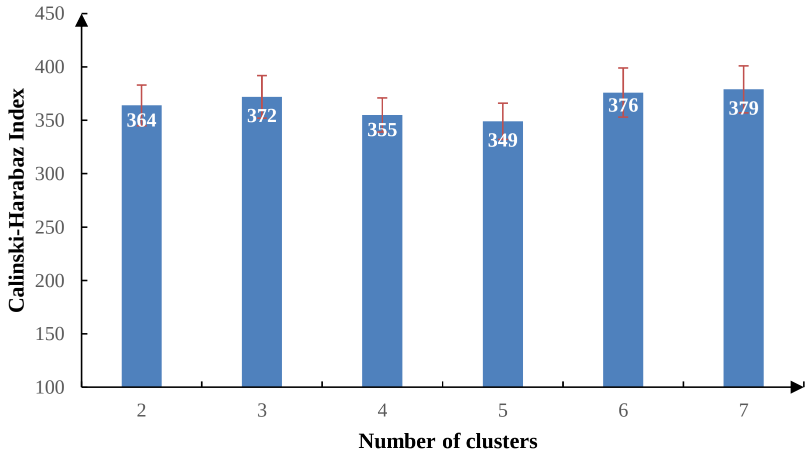

Fig. 5. The relationship between cluster numbers and Calinski-Harabaz Index based on the data for 105 cities from 2016 to 2018. The similar values of Calinski-Harabaz Index among different cluster numbers illustrate that the selection of cluster numbers does not significantly affect clustering performance (i.e. similar cities can be divided into same clusters).

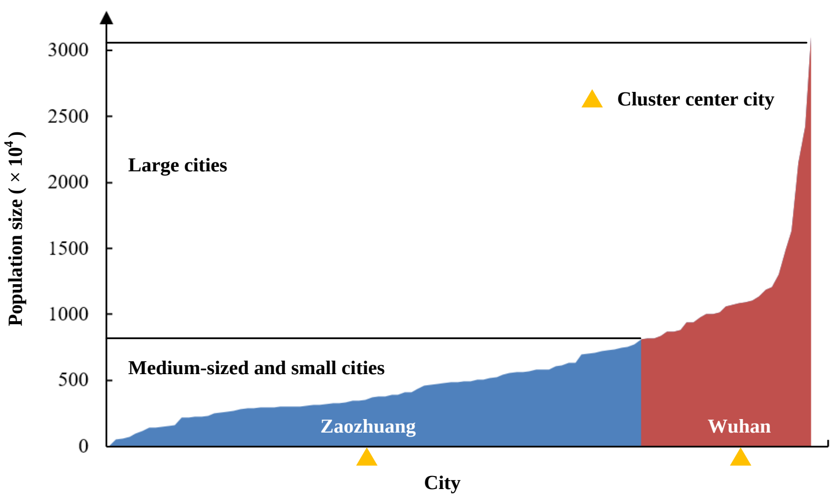

Fig. 6. Clustering results based on the data for 105 cities from 2016 to 2018. The cities are clustered in terms of population size because it produced the most significant effects on fire risk as shown in Table 2. The results show that the threshold between large cities and smaller cities and the cluster center cities of each cluster.

3.3 Multivariate regression models

Multivariate regression models were constructed for each city cluster to find the data governance determinants of fire risk. The models were trained using the sample data of 2016 and 2017. Then, the performance of the models was tested using the data of the year 2018.

3.3.1 Regression model for large cities

For large cities, the data governance determinants and their coefficients in the model are shown in Fig. 7. More specifically, Fig. 7(a) shows the results of the model which interprets fire frequency, and Fig. 7(b) shows the results of the model that interprets fire consequence. The figures show that the data governance determinants of fire frequency include: (1) infrastructure development, (2) data use, (3) high-value data sharing, (4) organizational maturity, and (5) data updating. Among them, infrastructure development most greatly affected fire frequency, which reflects that current intelligent technologies rely heavily on infrastructures and devices to provide basic data for fire risk analysis. With the support of multi-source fire risk data, the capabilities and ways of using these data are the most important task in data governance and have been shown a significant effect on reducing fire frequency. For example, the ‘Fire Eye’ system in Suzhou pre-evaluates the fire occurrence probabilities of various buildings (e.g., enterprises and institutions, residence, and public places such as schools and shopping centers) by analyzing the fire hazards, building structures, unsafe behaviors (e.g., excessive electricity consumption), and historical fire cases in those buildings based on multi-source big data and advanced AI technology. According to statistics, the accuracy of the ‘Fire Eye’ system for fire incident prediction has reached 70% [21]. The implication of high-value data sharing is to provide different kinds of data for more precise fire risk analysis. An instance is that the fire department in Atlanta achieved fire risk prediction with the integration of demographical data, building data, geographical data and fire incident data [20]. Organizational maturity is important in data governance as it promotes the coordination between departments. Although data governance has benefits for fire risk management, it is still an extra work for other departments, and thus an organizational promotion is needed. The effect of data updating was also significant as it improves the capacity of real-time fire risk analysis.

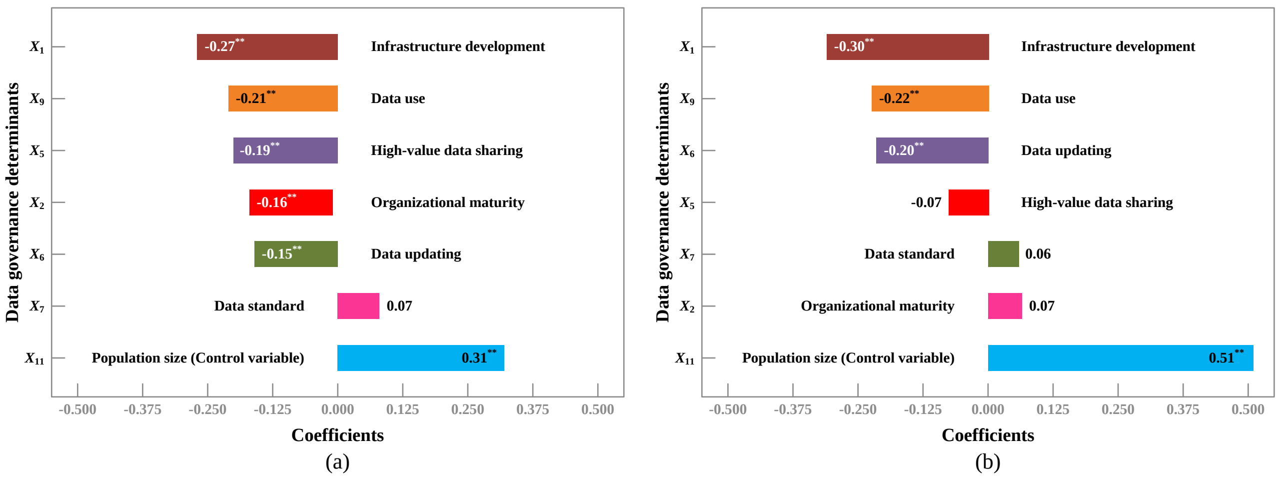

Fig. 7. Coefficients of data governance determinants in interpreting (a) fire frequency, and (b) fire consequence for large cities (based on the data for 30 cities of 2016 and 2017). The positive numbers indicate the corresponding determinant produces positive effects, and the negative ones indicate negative effects. The absolute values of the numbers indicate the strength of the effects. ** represents the coefficient is significant.

On the other hand, three data governance determinants are identified for interpreting fire consequences in large cities as shown in Fig. 7(b), i.e. (1) infrastructure development, (2) data use, and (3) data updating. Similarly to their effects on fire frequency, infrastructure development and data use improvement significantly facilitated reduction of fire consequence as they support effective collection and use of multi-source data for risk analysis, respectively. Data updating showed greater effects on fire consequence than frequency because fire risk treatment work (e.g., fire hazard treatment and firefighting) generally requires real-time fire risk recognition and guidance for operation. Compared to fire frequency, the effects of organizational maturity and data sharing on fire consequence were not significant partly because current big data technologies in fire risk treatment have not required much data from other departments. With the increasing development of novel technologies, the influences of the two factors are expected to show important effects.

The goodness of fit for large cities is shown in Fig. 8. In terms of fire frequency, the coefficient of determination (R^2) is 0.734 with training data, and that for the validation data is 0.682, which shows that the regression predictions can well approximate the real data values. The root-mean-square errors (RMSE) of the model for fire frequency are 0.09 and 0.06 (using training data and validation data), which illustrates good fitness from the perspective of prediction errors. The data fitness for fire consequence showed similar results.

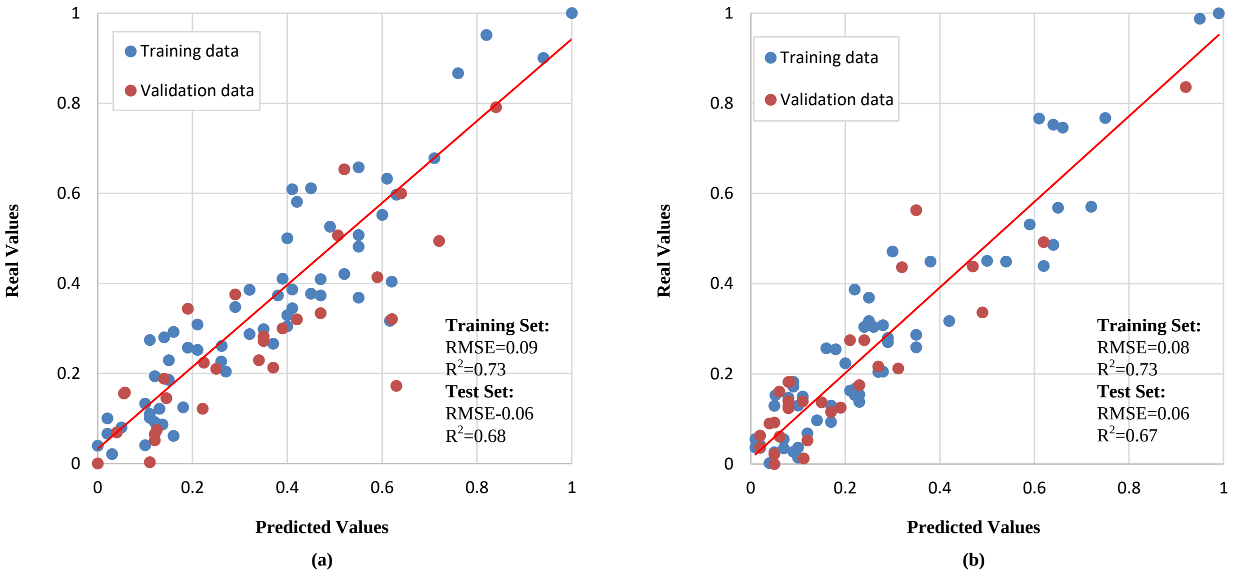

Fig. 8. Data fitting results of interpreting (a) fire frequency, and (b) fire consequence for large cities (based on the data for 105 cities from 2016 to 2018). The red lines are the fitting curves which use linear fitting functions. The coefficient of determination (R^2) and the root-mean-square errors (RMSE) that represent the goodness of fit are also displayed.

3.3.2 Regression model for medium-sized and small cities

Four data governance determinants are identified for reducing fire frequency in medium-sized and small cities as shown in Fig. 9(a), that is, (1) infrastructure development, (2) data use, (3) data updating, and (4) organizational maturity. Similarly to what was found for large cities, 'infrastructure development' and 'data use' were also proved important for reducing fire frequency in medium-sized and small cities. However, in smaller cities, 'data updating' had greater effects on the fire frequency than 'organizational maturity' had. This is partly because the organizational maturity in smaller cities is still low and thus its effects were still not significant. Generally, smaller cities have a smaller population and fewer fire incidents than large cities [16]. Hence, governments tend to not invest too much in big data technologies and data governance. Without the support of city governments, organizational coordination for data governance becomes difficult. This also explains why the influences of high-value data sharing and data standards were not significant.

Three data governance determinants ('infrastructure development', 'data use', and 'data updating') were identified for interpreting fire consequence reduction in medium-sized and small cities, as shown in Fig. 9(b). The results are similar to those in large cities. As analyzed in large cities, infrastructure development, 'data use' and 'data updating' help collect fire risk data, making use of these data and realizing real-time fire risk analysis, respectively. The higher the 'infrastructure development', 'data use' and 'data updating', the more precise and timelier the fire risk treatment. However, when compared with the results in large cities, 'data use' became more useful in fire consequence reduction, while the effects of 'infrastructure development' were reduced. The coefficients of other data governance determinants were not statistically significant.

The goodness of fit of the models for medium-sized and small cities is shown in Fig. 10. It can be seen that the data fitness was pretty good (with R^2 bigger than 0.65 and RMSE less than 0.10) even though it was worse than that for large cities.

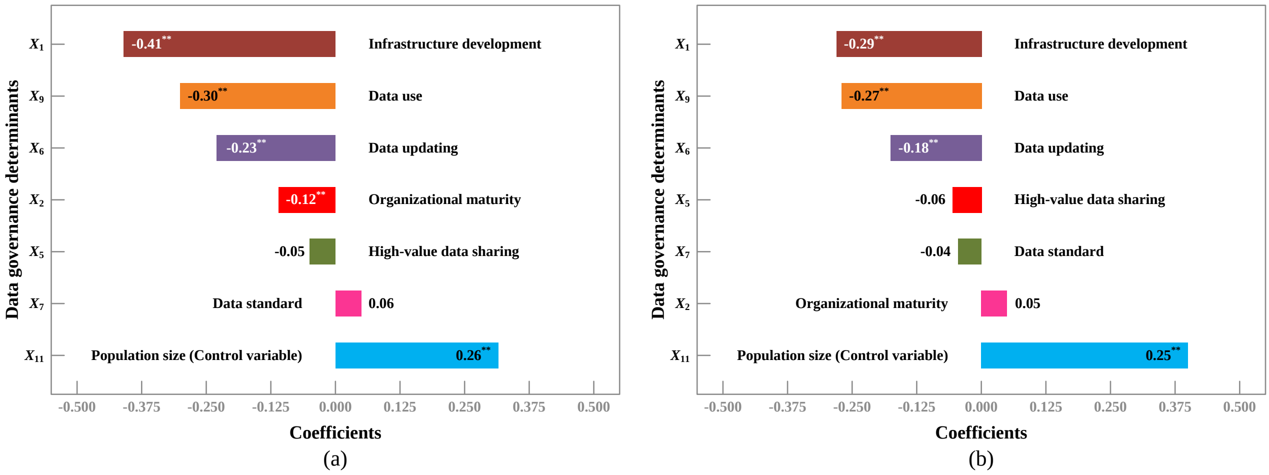

Fig. 9. Coefficients of data governance determinants in interpreting (a) fire frequency, and (b) fire consequence for medium-sized and small cities (based on the data for 30 cities of 2016 and 2017). The positive numbers indicate the corresponding determinant produces positive effects, and the negative ones indicate negative effects. The absolute values of the numbers indicate the strength of the effects. **indicates that the coefficient is significant.

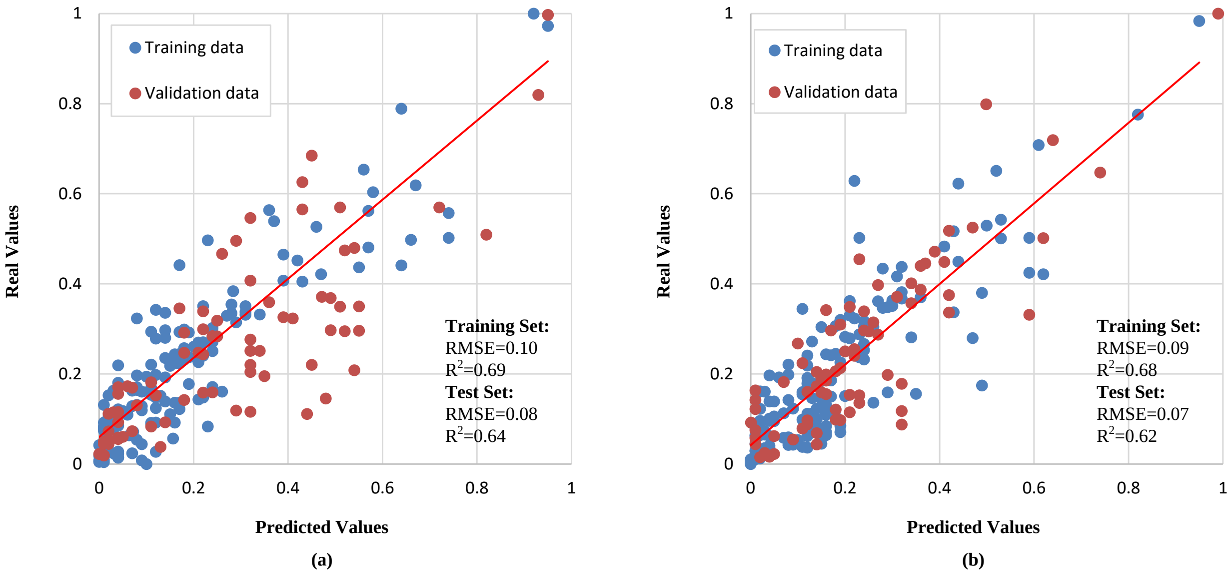

Fig. 10. Data fitting results of interpreting (a) fire frequency, and (b) fire consequence for medium-sized and small cities (based on the data for 105 cities from 2016 to 2018). The red lines are the fitting curves which use linear fitting functions. The coefficient of determination (R^2) and the root-mean-square errors (RMSE) that represent the goodness of fit are also displayed.

The results of the goodness of fit were also compared with previous studies (in terms of R^2), as shown in Table 4. It can be seen that the goodness of fit of the regression models in this paper performed better than that in existing studies, which illustrated the significance of incorporating data governance in understanding fire risk.

Table 4
Comparison of goodness of fit with other studies.

Indicator	Chhetri	Hastie	Hu	Large cities		Medium-sized and small cities	
	[60]	[16]	[4]	Fire frequency	Fire consequence	Fire frequency	Fire consequence
R^2	0.45	0.32	0.60	0.73 (0.68)	0.73 (0.67)	0.69 (0.64)	0.68 (0.62)

Note: the goodness of fit to the training data is displayed in the brackets, and that to the validation data is outside the brackets.

3.4 Extensive clustering analysis of the data governance indicators

To comprehensively analyze the effects of data governance on urban fire risk, a wide range of indicators have been included and validated in this paper. However, in empirical research, the reliability of the results is likely to be limited by the indicator diversity. Although the effects of each indicator have been analyzed, meaningful empirical research should also provide an in-depth quantitative analysis for explaining the results [61]. In this part, a clustering analysis of the various original indicators and data was performed to provide a deeper understanding of the abovementioned statistical results. Specifically, a hierarchical clustering method was selected for grouping the indicators as it does not require to pre-specify the number of clusters [45]. Similar to the city clustering approach, the clustering performance (measured by the Calinski-Harabaz Index) among different cluster numbers was compared to obtain the optimal clusters. Finally, three indicator clusters and their members were obtained. The clustering results are shown in Table 5.

Table 5

The clustering results of the data governance indicators. The members that are included in each indicator cluster are presented.

Data governance indicators	Clustering results		
	Indicator cluster 1	Indicator cluster 2	Indicator cluster 3
Infrastructure development	Included	/	/
Organizational maturity	Included	/	/
Policy support	Included	/	/
Data sharing	/	Included	/
High-value data sharing	/	Included	/
Data updating	/	Included	/
Data standard	Included	/	/
Data use: Data sources	/	/	Included
Data use: Data use level	/	/	Included
Data security	/	Included	/

It can be seen in Table 5 that each indicator cluster has its own features. The members of indicator cluster 1 include infrastructure development, organizational maturity, policy support, and data standard. These indicators mainly show that data governance is a complex work and requires the support of various aspects for effective data collection, management, and use. Cluster 2 includes (high-value) data sharing, data updating, and data security. It can be seen that these data activities mainly reflect the high-quality data management work for supporting fire services. In data governance, those work is vital to realizing high-quality data supply. Finally, cluster 3 focuses on the effective use of multi-source data. The members reflect the direct data support for fire risk management. Consequently, the three clusters can be named as data governance support, high-quality data management, and fire service data use, respectively. The ten broad indicators were finally converted to three meaningful factors.

The three indicator clusters were then used for explaining the data governance effects on fire risk. Combined with the multivariate regression results, it can be seen that the data governance determinants cover the members of all the three clusters, which indicates that those factors jointly facilitate data-driven fire risk management and thus reduce urban fire risk levels. This can be explained by the existing data governance mechanisms in public safety fields [29]. Specifically, the data governance support provides the initial environment for big data value creation and includes various factors such as infrastructure development and policy support. Under the initial environment, various data management work (e.g., data sharing and data security) was adopted to provide high-quality data to fire departments for improving fire risk management. Finally, the big data value in fire risk management is discovered through the effective use of multi-source data in fire safety tasks (as shown in Table 1). Through the multivariate regression analysis in this paper, the key members of the three clusters have been validated, e.g., the infrastructure development in data governance support, and the high-value data sharing in data management.

4 Discussion

4.1 Applicability

Fire risk assessment, particularly if it is to be quantitative, is challenging and has both physical and social determinants [4, 14]. From social perspective, data governance serves the use of big data in fire risk management and can help reduce fire risk to some extent. However, the fire risk situation of different cities varies and data governance cannot completely eliminates the threats of fires. Consequently, understanding the applicability of data governance statistics is important for using these statistics. Currently, the data governance statistics mainly benefit two types of cities regarding effective fire risk management, i.e. cities with high urbanization and cities with good fire prevention.

First, in cities with high urbanization (e.g. Singapore city), there are generally dense and numerous buildings, which produce various kinds of data (e.g. building structure, population size, electricity, and water consumption) for comprehensively analyzing fire risks. As buildings are the main places where fire incidents occur, it is reasonable to adopt data governance measures in these cities to make full use of these data [14]. This illustrates why successful cases of intelligent fire risk management basically emerge in cities or areas with high urbanization, e.g. the 'Firebird' system in Atlanta, US, the 'Fire Eye' system in Suzhou, China and the fire prediction platform in London. In contrast, in cities with low urbanization, there are relatively fewer and sparser buildings, and thus the overall levels of fire risk in these cities are lower [4, 16]. Hence, the demands for intelligent fire risk management in these cities are less strong and the governments in these cities tend to not invest too much in data governance. This applicability is also supported by the statistical results, i.e., more data governance indicators in large cities produced effects on fire risk than those in smaller cities. However, with the development of urbanization in the smaller cities, the demands for more effective fire risk management will inevitably increase and thus the data governance statistics are expected to create benefits in the near future.

Second, in cities with good fire prevention, generally fewer incidents are caused by indoor human behaviors [62], e.g. playing with fire, smoking, and the misuse of equipment or appliances, which intelligent technologies are hard to deal with due to the difficulties of collecting relevant data. Hence, governments can pay more attention to the fire threats related to many other features, e.g. building structures, socio-economic and demographic characteristics, crimes, and deprivation. Further, the analysis of these data facilitates the use of big data technologies and relevant data governance [54], and thus the data governance statistics become more useful. For example, the fire risk levels of some US cities were found to be highly related to weather conditions, and thus data governance can benefit the fire risk management in these cities as weather-related data is easy to be collected [13]. However, for other cities where fire risk is highly related to indoor human behaviors, the data governance of weather-related data is obviously useless. It should be noted that with the continuous development of big data technologies, data governance can benefit more and more cities. A typical instance is that the fire department in Guizhou is trying to analyze residents' indoor behavior using electricity data [63].

4.2 Implications

The reported findings extend the research on urban fire risk by explaining it from a social perspective, i.e. the effects of data governance (as a means of fire department intervention) on fire risk management. In previous studies, fire risk has been interpreted by many socio-economic features, e.g. population size, industrial development, and income [4]. However, these socio-economic features often reflect the behaviors of building occupants instead of the intervention of fire departments. Hence, the discovery and adaptation of effective measures for urban fire risk management have been greatly limited. In particular, the data governance endeavors for the use of big data in fire risk management have not been well analyzed. This paper presents an illustration of data governance analysis based on China's experiences. By noting that data governance factors (e.g. infrastructure development, data use, and high-quality data sharing) significantly reduces fire risk, interpretation of the effects of data governance on fire risk is finally reached, i.e. data governance can benefit fire risk management and thus reduce fire risk through the effective collection, management, and use of big data.

Second, the data governance statistics can be used for improving fire risk management of a wide range of cities, especially those with high urbanization and those with good fire prevention. Current fire risk management is usually analyzed and improved using traditional indicators (which have little relevance with fire technologies), e.g. statistics of fire stations, fire vehicles, and home fire safety visits [18]. The indicators should be updated as the emerging big data technologies are playing more and more important roles in fire risk management. As presented in this paper, the data governance determinants are highly correlated with fire risk and thus can be selected as new indicators for evaluating the performance of fire risk management.

Third, this study extends data governance research to a fire risk management context. Existing research on data governance tends to show the potential value of specific big data technologies but overlooks other data governance work (e.g. data sharing, data quality control, and organizational cooperation) [31, 32]. Some studies investigated the effects of data governance comprehensively but lacked a quantitative verification of the conclusions [33]. This paper reduces the gap by examining the effects of a wide range of data governance characteristics on fire risk through a statistical analysis of 105 cities in China.

Fourth, guidelines for urban fire risk management for similar jurisdictions around the world as well as Chinese cities are also provided. The solution has been proved to be effective based on data of more than 100 cities in China and thus have relatively high generality. For the jurisdictions that have similar fire risk situations with China, the city governments can refer to the following conclusions for adopting reasonable data governance measures to improve fire risk management.

5 Conclusion

This study investigated the relationship between data governance conditions and urban fire risk by constructing multivariate regression models based on data for 105 Chinese cities from 2016 to 2018. The empirical results showed the following conclusions:

First, six data governance determinants of urban fire risk were identified through Pearson correlation analysis, i.e. infrastructure development, organizational maturity, high-value data sharing, data updating, data standard, and data use. In addition, the effects of major socio-economic on fire risk were also validated in the correlation analysis with population size selected as a control variable in the constructed multivariate regression models.

Second, cities were clustered into two clusters with population size selected as the clustering indicator, i.e. large cities (with a higher population) and medium-sized and small cities (with lower population). Regression models of the two city clusters were then constructed with R^2 exceeding 0.7. The regression results indicated that most of the representative cities have low levels of population, with only 30 of the 105 sample cities having higher population.

Third, the influences of data governance on fire risk in the two types of cities were analyzed in detail. In large cities, infrastructure development, data use, high-value data sharing, organizational maturity, and data updating were found to have significant negative effects on fire incidence while population size significantly increased the incident. In smaller cities, fire frequency had significant negative relations with infrastructure development, data use, data updating, and organizational maturity. Compared with those on fire frequency, the effects of current data governance on fire consequence (measured by fire direct property losses) were smaller in both types of cities, with infrastructure development, data use, and data updating significantly related to it.

Fourth, the data governance indicators were grouped to further illustrate the data governance effects on urban fire risk. Through a hierarchical clustering analysis, the effects of data governance were explained from three key factors, i.e., data governance support, high-quality data management, and fire service data use.

There are still some limitations in this study. Firstly, the sample size is relatively small compared with the total number of cities in China, and thus cross-city discrepancies might exist. With the increasing use of data governance measures, future research should take care of the data from more cities to achieve refined findings. Secondly, the sample in this study is mainly composed of provincial and prefecture-level cities. In fact, county-level cities and communities are playing a more and more important role in fire risk management, and thus future studies should pay more attention to the data governance in those cities with lower administrative levels. Thirdly, fire prevention (focusing on educating the public to take precautions to prevent potential fires) is a crucial factor that moderates the influences of data governance, which lacks relevant statistical support in this paper as the statistical data of fire prevention is scarce. Hence, future studies should investigate more quantitative indicators to prove the fire prevention effects. Fourthly, this study assumes that within a year, the data governance conditions are continuously changed and their effects are rapidly reflected in annual fire statistics. However, that may benefit from the efficient and large-scale data governance projects in China. Future studies can focus on finding out the precise mechanism of how the effects change over time in different regions. Finally, the effects of data governance on detailed fire risk management tasks (e.g., the emergency responses of fire command centers) deserve in-depth statistical analysis, and there should be sufficient data for further discussion in the future.

Acknowledgements

This work is supported by the Major Research Project of Nation Natural Science Foundation of China named "Big data Driven Management and Decision-making Research" (No. 91746207), the General Program of Nation Natural Science Foundation of China (No. 71774043) and the Fundamental Research Funds for the Central Universities (No. 20720221020).

References

[1] James SL, Lucchesi LR, Bisignano C, Castle CD, Dingels ZV, Fox JT, et al., Epidemiology of injuries from fire, heat and hot substances: global, regional and national morbidity and mortality estimates from the Global Burden of Disease 2017 study, Injury Prevention, 2020;26: i36-i45.

[2] International Association of Fire and Rescue Services, World fire statistics. https://www.ctif.org/world-fire-statistics, 2020 (accessed 21 October 2021)

[3] U.S. Fire Administration, U.S. fire statistics. https://www.usfa.fema.gov/data/statistics/, 2021 (accessed 23 October 2021)

[4] Hu J, Shu XM, Xie ST, Tang SY, Wu JJ, Deng BY, Socioeconomic determinants of urban fire risk: A city-wide analysis of 283 Chinese cities from 2013 to 2016, Fire Safety Journal, 2019;110.

[5] The State Council of China, China sees 13.6% drop in deaths from fire in 2020. http://english.www.gov.cn/statecouncil/ministries/202101/22/content_WS600ac5b3c6d0f725769445a9.html, 2021 (accessed 23 October 2021)

[6] Nishino T, Tanaka T, Hokugo A, An evaluation method for the urban post-earthquake fire risk considering multiple scenarios of fire spread and evacuation, Fire Safety Journal, 2012;54: 167-80.

[7] Ding L, Khan F, Ji J, Risk-based safety measure allocation to prevent and mitigate storage fire hazards, Process Safety and Environmental Protection, 2020;135: 282-93.

[8] Liu ZG, Li XY, Jomaas G, Identifying Community Fire Hazards from Citizen Communication by Applying Transfer Learning and Machine Learning Techniques, Fire Technology, 2021;57: 2809-38.

[9] Baquedano Juliá P, Ferreira TM, Rodrigues H, Post-earthquake fire risk assessment of historic urban areas: A scenario-based analysis applied to the Historic City Centre of Leiria, Portugal, International Journal of Disaster Risk Reduction, 2021;60: 102287.

[10] Kwegyir-Afful E, Effects of an engaging maintenance task on fire evacuation delays and presence in virtual reality, International Journal of Disaster Risk Reduction, 2021: 102681.

[11] Xin J, Huang CF, Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management, Fire Safety Journal, 2013;62: 72-78.

[12] Guldaker N, Hallin PO, Spatio-temporal patterns of intentional fires, social stress and socio-economic

determinants: A case study of Malmo, Sweden, Fire Safety Journal, 2014;70: 71-80.

[13] Agarwal P, Tang JL, Narayanan ANL, Zhuang J, Big Data and Predictive Analytics in Fire Risk Using Weather Data, Risk Analysis, 2020;40: 1438-49.

[14] Liu F, Zhao SZ, Weng MC, Liu YQ, Fire risk assessment for large-scale commercial buildings based on structure entropy weight method, Safety Science, 2017;94: 26-40.

[15] Jennings CR, Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature, Fire Safety Journal, 2013;62: 13-19.

[16] Hastie C, Searle R, Socio-economic and demographic predictors of accidental dwelling fire rates, Fire Safety Journal, 2016;84: 50-56.

[17] Corcoran J, Higgs G, Rohde D, Chhetri P, Investigating the association between weather conditions, calendar events and socio-economic patterns with trends in fire incidence: an Australian case study, Journal of Geographical Systems, 2011;13: 193-226.

[18] Sund B, Bonander C, Jakobsson N, Jaldell H, Do home fire and safety checks by on-duty firefighters decrease the number of fires? Quasi-experimental evidence from Southern Sweden, Journal of Safety Research, 2019;70: 39-47.

[19] Taylor M, Oakford G, Appleton D, Fielding J, Fire Prevention Targeting by Merseyside Fire and Rescue Service in UK, Fire Technology, 2022.

[20] Madaio M, Chen ST, Haimson OL, Zhang WW, Cheng X, Hinds-Aldrich M, et al. Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. In. Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. San Francisco, CA, 2016, pp. 185-94.

[21] Xinhua Daily, 'Fire Eye System' in Suzhou has largely improved fire prevention and control precision. http://js.cri.cn/chinanews/20171206/8553eba1-8664-63d2-bb3c-c0535f4255d2.html, 2017 (accessed 23 October 2021)

[22] Clark N, Guiffault F, Seeing through the clouds: Processes and challenges for sharing geospatial data for disaster management in Haiti, International Journal of Disaster Risk Reduction, 2018;28: 258-70.

[23] Araujo Lima GP, Viana Barbosa JD, Beal VE, Moret S. Gonçalves MA, Souza Machado BA, Gerber JZ, et al., Exploratory analysis of fire statistical data and prospective study applied to security and protection systems, International Journal of Disaster Risk Reduction, 2021;61: 102308.

[24] Liu J, Li TR, Xie P, Du SD, Teng F, Yang X, Urban big data fusion based on deep learning: An overview, Information Fusion, 2020;53: 123-33.

[25] Goel P, Datta A, Mannan MS. Application of Big Data analytics in process safety and risk management. In. Application of Big Data analytics in process safety and risk management. Boston, MA, 2017, pp. 1143-52.

[26] Thompson N, Ravindran R, Nicosia S, Government data does not mean data governance: Lessons learned from a public sector application audit, Government Information Quarterly, 2015;32: 316-22.

[27] Sunil S, Big data governance, Information Asset, LLC, 2012.

[28] Wikipedia, Data governance. https://en.wikipedia.org/wiki/Data_governance, 2022 (accessed 25 April 2022)

[29] Liu Z-g, Li X-y, Zhu X-h, Scenario modeling for government big data governance decision-making: Chinese experience with public safety services, Information & Management, 2022;59: 103622.

[30] Euronews, AP, Firefighters in China are using big data to predict fires before they happen. https://www.euronews.com/next/2021/04/29/firefighters-in-china-are-using-big-data-to-predict-fires-before-they-happen, 2021 (accessed 21 April 2022)

[31] Wu XQ, Park Y, Li A, Huang XY, Xiao F, Usmani A, Smart Detection of Fire Source in Tunnel Based on the Numerical Database and Artificial Intelligence, Fire Technology, 2021;57: 657-82.

[32] Granda S, Ferreira TM, Assessing Vulnerability and Fire Risk in Old Urban Areas: Application to the Historical Centre of GuimarAes, Fire Technology, 2019;55: 105-27.

[33] Cumbie BA, Sankar CS, Choice of governance mechanisms to promote information sharing via boundary objects in the disaster recovery process, Information Systems Frontiers, 2012;14: 1079-94.
[34] UNISDR, Terminology on disaster risk reduction. https://www.unisdr.org/we/inform/terminology, 2017 (accessed 25 January 2022)
[35] Djoumessi YF, Eyike Mbongo LdB, An analysis of information Communication Technologies for natural disaster management in Africa, International Journal of Disaster Risk Reduction, 2022;68: 102722.
[36] Ahsan MN, Khatun A, Fostering disaster preparedness through community radio in cyclone-prone coastal Bangladesh, International Journal of Disaster Risk Reduction, 2020;49: 101752.
[37] Dinh NC, Ubukata F, Tan NQ, Ha VH, How do social connections accelerate post-flood recovery? Insights from a survey of rural households in central Vietnam, International Journal of Disaster Risk Reduction, 2021;61: 102342.
[38] China Information Technology Industry Federation, China's governmental data governance development report. https://www.sohu.com/a/414591745_100039689, 2020 (accessed 23 October 2021)
[39] Lab for Digital & Mobile Governance, 2016 China local government data opening report. www.dmg.fudan.edu.cn, 2016 (accessed 25 October 2021)
[40] Lab for Digital & Mobile Governance, 2017 China local government data opening report. www.dmg.fudan.edu.cn, 2017 (accessed 25 October 2021)
[41] Lab for Digital & Mobile Governance, 2018 China local government data opening report. www.dmg.fudan.edu.cn, 2018 (accessed 25 October 2021)
[42] Abraham R, Schneider J, vom Brocke J, Data governance: A conceptual framework, structured review, and research agenda, International Journal of Information Management, 2019;49: 424-38.
[43] Al-Ruithe M, Benkhelifa E. Cloud Data Governance Maturity Model. In. Cloud Data Governance Maturity Model. Univ Cambridge, Churchill Coll, Cambridge, ENGLAND, 2017.
[44] Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly, 2020;37.
[45] Xu Q, Zhang Q, Liu JP, Luo B, Efficient synthetical clustering validity indexes for hierarchical clustering, Expert Systems with Applications, 2020;151.
[46] Fire Service Bureau, 2016 China Fire Yearbook, Yunnan Personnel Press, Kunming, 2017 (in Chinese)
[47] Fire Service Bureau, 2017 China Fire Yearbook, Yunnan Personnel Press, Kunming, 2018 (in Chinese)
[48] Fire Service Bureau, 2018 China Fire Yearbook, Yunnan Personnel Press, Kunming, 2019 (in Chinese)
[49] Thompson GL, An SPSS implementation of the nonrecursive outlier deletion procedure with shiftingz score criterion (Van Selst & Jolicoeur, 1994), Behavior Research Methods, 2006;38: 344-52.
[50] National Bureau of Statistics of China, 2016 China City Statistical Yearbook, China Statistics Press, Beijing, 2017 (in Chinese)
[51] National Bureau of Statistics of China, 2017 China City Statistical Yearbook, China Statistics Press, Beijing, 2018 (in Chinese)
[52] National Bureau of Statistics of China, 2018 China City Statistical Yearbook, China Statistics Press, Beijing, 2019 (in Chinese)
[53] Rymes MD, Myers DR, Mean preserving algorithm for smoothly interpolating averaged data, Solar Energy, 2001;71: 225-31.
[54] Benfeldt O, Persson JS, Madsen S, Data Governance as a Collective Action Problem, Information Systems Frontiers, 2020;22: 299-313.
[55] Malik P, Governing Big Data: Principles and practices, Ibm Journal of Research and Development, 2013;57.
[56] Silva MM, Poleto T, Silva LCE, de Gusmao APH, Costa A, A Grey Theory Based Approach to Big Data Risk Management Using FMEA, Mathematical Problems in Engineering, 2016;2016.

[57] Alhassan I, Sammon D, Daly M, Data governance activities: an analysis of the literature, Journal of Decision Systems, 2016;25: 64-75.
[58] Zhang YK, Li YL, Song JZ, Chen XL, Lu Y, Wang WK, Pearson correlation coefficient of current derivatives based pilot protection scheme for long-distance LCC-HVDC transmission lines, International Journal of Electrical Power & Energy Systems, 2020;116.
[59] Petrella L, Raponi V, Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress, Journal of Multivariate Analysis, 2019;173: 70-84.
[60] Chhetri P, Corcoran J, Stimson RJ, Inbakaran R, Modelling potential Socio‐economic determinants of building fires in south east Queensland, Geographical Research, 2010;48: 75-85.
[61] Nicholson D, Vanli OA, Jung S, Ozguven EE, A spatial regression and clustering method for developing place-specific social vulnerability indices using census and social media data, International Journal of Disaster Risk Reduction, 2019;38: 101224.
[62] Sufianto H, Green AR, Urban Fire Situation in Indonesia, Fire Technology, 2012;48: 367-87.
[63] Zhu J, Cheng Y, Smart use of big data to prevent electrical fires. http://www.cbdio.com/BigData/2018-09/21/content_5843943.htm, 2018 (accessed 21 October 2021)

Appendix: List of open data and governmental official websites

Table 5 List of open data and governmental official websites

Province	City	Website type	Internet address / URL
Shandong	Jinan	Open data platform	http://data.jinan.gov.cn/
	Qingdao	Open data platform	http://data.qingdao.gov.cn/odweb/catalog/index.htm
	Yantai	Open data platform	ytdata.sd.gov.cn/
	Weihai	Open data platform	whdata.sd.gov.cn/
	Weifang	Open data platform	http://wfdata.sd.gov.cn/
	Rizhao	Open data platform	rzdata.sd.gov.cn/
	Binzhou	Open data platform	bzdata.sd.gov.cn/
	Zaozhuang	Open data platform	www.zzdata.gov.cn/
	Dongying	Open data platform	dydata.sd.gov.cn/
	Jining	Open data platform	jindata.sd.gov.cn/
	Linyi	Open data platform	lydata.sd.gov.cn/
	Dezhou	Open data platform	dzdata.sd.gov.cn/
Heilongjiang	Harbin	Open data platform	http://data.harbin.gov.cn/
Hubei	Wuhan	Open data platform	http://www.wuhandata.gov.cn/
	Jingmen	Government website	http://dh.jingmen.gov.cn/col/col876/index.html
	Qianjiang	Government website	http://www.hbqj.gov.cn/xxgk/xxgkml/szfxxgkml/sjfb/
	Huanggang	Government website	http://www.hg.gov.cn/col/col7161/
	Xiangyang	Government website	www.xf.gov.cn/
Anhui	Anqing	Government website	http://aqxxgk.anqing.gov.cn/index_bm.php?unit=HA110
	Bozhou	Government website	http://sjzyj.bozhou.gov.cn/
	Chizhou	Government website	http://chizhou.gov.cn/DataRelease/
	Chuzhou	Government website	http://sjzyj.chuzhou.gov.cn/
	Fuyang	Open data platform	m.fy.gov.cn/openData/
	Huainan	Government website	http://sjzyj.huainan.gov.cn/
	Huangshan	Open data platform	http://www.huangshan.gov.cn/DataDevelopment/
	Lu’an	Open data platform	http://data.luan.gov.cn:8081/dop/
	Maanshan	Open data platform	www.mas.gov.cn/content/column/4697374
	Wuhu	Open data platform	https://data.wuhu.cn/
Beijing	Beijing	Open data platform	https://data.beijing.gov.cn/
Fujian	Fuzhou	Open data platform	http://data.fuzhou.gov.cn/
	Quanzhou	Open data platform	http://www.quanzhou.gov.cn/zfb/xxgk/
	Xiamen	Open data platform	https://data.xm.gov.cn/opendata/
Gansu	Lanzhou	Government website	http://dsjj.lanzhou.gov.cn/
Guangdong	Chaozhou	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/515
	Foshan	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/38
	Guangzhou	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/27
	Heyuan	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/510
	Huizhou	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/30
	Jiangmen	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/47
	Jieyang	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/516
	Maoming	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/31
	Meizhou	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/58

	Shantou	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/28
	Shanwei	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/59
	Shaoguan	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/37
	Shenzhen	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/29
	Yangjiang	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/511
	Zhanjiang	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/32
	Zhongshan	Open data platform	https://gddata.gd.gov.cn/data/dataSet/toDataSet/dept/514
Guangxi	Guilin	Government website	http://www.guilin.gov.cn/glsj/sjkf/
	Nanning	Government website	http://www.nanning.gov.cn/sjfw/sjkf/
Guizhou	Guiyang	Open data platform	https://data.guiyang.gov.cn/city/index.htm
	Liupanshui	Open data platform	http://data.gzlps.gov.cn/
	Tongren	Open data platform	http://www.gztrdata.gov.cn/index.html
Hainan	Sanya	Open data platform	http://dataopen1.sanya.gov.cn/
	Wuzhishan	Government website	http://wzs.hainan.gov.cn/wzs/0900/list2.shtml
Hebei	Shijiazhuang	Open data platform	http://opendata.sjz.gov.cn/portal/public/home
Henan	Hebi	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4106
	Jiyuan	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4190
	Luohe	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4111
	Nanyang	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4113
	Pingdingshan	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4104
	Sanmenxia	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4112
	Xinxiang	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4107
	Zhengzhou	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4101
	Zhoukou	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4116
	Zhumadian	Open data platform	http://data.hnzwfw.gov.cn/odweb/catalog/index.htm?region_code=4117
Hunan	Xiangtan	Open data platform	http://www.xiangtan.gov.cn/1029/index.htm
	Yueyang	Open data platform	http://www.yueyang.gov.cn/yytj/default.htm
Jilin	Changchun	Government website	http://tjj.changchun.gov.cn/tjsj/
Jiangsu	Nantong	Open data platform	http://data.nantong.gov.cn/home/index.html
	Suzhou	Open data platform	http://www.suzhou.gov.cn/OpenResourceWeb/resources?isAsc=false
	Xuzhou	Open data platform	http://www.xz.gov.cn/zgxz/sjkf/
	Zhenjiang	Government website	http://tjj.zhenjiang.gov.cn/
Jiangxi	Jiujiang	Open data platform	https://www.jiujiang.gov.cn/sjkf/
	Jian	Government website	http://www.jian.gov.cn/xxgk.html
	Nanchang	Open data platform	http://www.nc.gov.cn/ncszf/sjkfn/sjkfej.shtml
Liaoning	Dalian	Government website	http://www.dl.gov.cn/xxgk.vm?lid=3
Nei Mongol	Hohhot	Open data platform	http://www.huhhot.gov.cn/zfsj/
Ningxia	Yinchuan	Open data platform	http://data.yinchuan.gov.cn/odweb/
Shanxi	Jincheng	Government website	http://www.jcgov.gov.cn/?dw=zwgk
Shaanxi	Hanzhong	Open data platform	http://en.hanzhong.gov.cn/xxgk/
	Xi’an	Government website	http://tjj.xa.gov.cn/tjsj/1.html
Shanghai	Shanghai	Open data platform	https://data.sh.gov.cn/
	Pudong	Open data platform	https://data.sh.gov.cn/view/data-resource/index.html
Sichuan	Chengdu	Open data platform	http://www.cddata.gov.cn/oportal/index
	Deyang	Government website	http://www.deyang.gov.cn/info/iList.jsp?tm_id=42

	Guang'an	Open data platform	http://www.guang-an.gov.cn/gasrmzfw/shujkfpt/sjkfpt_index.shtml
	Leshan	Open data platform	http://data.leshan.gov.cn/
	Meishan	Open data platform	http://www.ms.gov.cn/ index.htm
	Neijiang	Open data platform	http://data.neijiang.gov.cn/
	Ya'an	Open data platform	http://data.yaan.gov.cn/index/index.html
	Ziyang	Government website	http://gk.ziyang.gov.cn/
	Zigong	Open data platform	http://www.zg.gov.cn/sjkf-new
Xizang	Lasa	Government website	www.lasa.gov.cn/
Zhejiang	Hangzhou	Open data platform	https://data.hz.zjzwfw.gov.cn/
	Huzhou	Open data platform	http://data.huz.zjzwfw.gov.cn/
	Jiaxing	Open data platform	http://data.zjzwfw.gov.cn/jdop_front/index.do
	Jinhua	Open data platform	http://data.jh.zjzwfw.gov.cn/
	Ningbo	Open data platform	http://data.nb.zjzwfw.gov.cn/nbdata/fore/index.html
	Quzhou	Open data platform	http://data.qz.zjzwfw.gov.cn/jdop_front/index.do
	Shaoxing	Open data platform	https://data.sx.zjzwfw.gov.cn/
	Taizhou	Open data platform	http://data.taz.zjzwfw.gov.cn/tz/
	Zhoushan	Open data platform	http://data.zs.zjzwfw.gov.cn/
Chongqing	Chongqing	Open data platform	http://data.tjj.cq.gov.cn/govindex.htm