Article

Evaluation on Sports Facility Resource in Primary School Using a Combined Approach of Unsupervised Machine Learning: a Case Study of Shanghai, China

Jun Xia [1], Pei-Jie Chen [2], *, Ji-Hong Wang [3], Jie Zhuang [4], Zhen-Bo Cao [5], Qiang Zhang [6]

[1] School of Kinesiology, Shanghai University of Sport, Shanghai, China; dx00122@163.com

[2] School of Kinesiology, Shanghai University of Sport, Shanghai, China; chenpeijie@sus.edu.cn

[3] School of Kinesiology, Shanghai University of Sport, Shanghai, China; wjh@sus.edu.cn

[4] School of Kinesiology, Shanghai University of Sport, Shanghai, China; zhuangjiesh@163.com

[5] School of Kinesiology, Shanghai University of Sport, Shanghai, China; caozhenbo@sus.edu.cn

[6] School of Mechanical and Power Engineering, Tongji University, Shanghai, China; zhangqiang8819@126.com

* Correspondence: chenpeijie@sus.edu.cn; Tel.: +86-21-65508039; Fax.: +86-21-65508050

Abstract: The aim of this study is (a) to develop, test, and employ a combined method of unsupervised machine learning to objectively assess the condition of sports facility in primary schools (PSSFC) and (b) examine the geographical and typological association with PSSFC. Based on the Sixth National Sports Facility Census (NSFC), six PSSFC indicators (indoor and outdoor facility included) were selected as the measurements and decomposed by using the t-stochastic neighbor embedding (t-SNE). Thereafter, the Fuzzy C-mean (FCM) algorithm was used to cluster the same type of PSSFC with selecting the optimum numbers of evaluation level. Overall 845 primary schools in Shanghai, China were recruited and tested by this combined approach of unsupervised machine learning. In addition, the two-way analysis of covariance was used to examine the location and types of school associated with PSSFC variables in each level. The combined method was found to have acceptable reliability and good interpretability, differentiating PSSFC into five gradient levels. The characteristics of PSSFC differ by the location and school type of individual school. Our findings are conducive to the regionalized and personalized intervention and promotion on the children's physical activity (PA) upon the practical situation of particular schools.

Keywords: School Sports Facility; Assessment; T-SNE; Fuzzy C mean; Unsupervised Learning

1. Introduction

The prosperity of a country, to some extent, relies on the forged productive manners of young generation, among which physical fitness (PF) is regarded as the fundamental and vital element [1]. Sufficiency of physical activity at moderate to vigorous level is well-researched contribution to facilitate the health-related PF and reduction of childhood obesity, non-communicable chronic diseases, heart disease and diabetes [2-6]. The World Health Organization (WHO) recommends children and adolescence aged from 5 to 17 years old to participate moderate-to-vigorous intensity physical activity (MVPA) for one hour per day[7] .Notwithstanding the benefits of PA, the global average grade remains D level on 9 indicators of PA, announced by the Global Matrix 2.0 [8]. In particular, sedentary behaviors and organized sports participation in China rated with F level presents the sequential recession of PA levels [9-11] and growth of sedentary behavior among school-age population [12, 13].

In general, the schoolchildren's PF and PA are diversified not only by the individual characteristics [14] but also by the environmental factors [15-17]. Among the latter, in contrast with the family and community, the school is widely perceived as the subject of admission and responsibility on the promotion and intervention of schoolchildren's PA [18]. The previous investigators have identified the positive correlation between the diversity of school sports facilities (SSF) and children's PA [19-21]. Moreover, in addition to supporting the high quality of physical education lessons, the excellence of SSF motivates schoolchildren to be physical active during the recess and lunch time [22]. The discrepancy on SSF determines the variance of schoolchildren's accessibility and availability on the school-based PA participation. Therefore, the condition of SSF can be regarded as an indirect and non-ignorable factor of PA promotion in childhood.

In recent years, a number of researchers have sought to classify SSF, aiming to quantify the correlation of SSF differentiation with students' PA and PF. Fórnias *et al.* [23] introduced the "number of PA facilities" variable into analysis the relationship of adolescent PA and extracurricular sports activities. This variable consisted of the quantities and using condition of 6 types of SSF, and was analyzed its proportion of schools. Lo *et al.* [24] conducted a cross-sectional study on the Taiwanese adolescent school sports condition and relevant extracurricular activities, with the consideration of the investigated schools with or without the school sports field and gymnasium. Haug *et al.* [25] generated outdoor facility index (summarized and standardized score of the outdoor SSF numbers), based on a school-level questionnaire for Norwegian primary and secondary schools. The results revealed its correlation with students' physical activity. Bevans *et al.* [26] proposed a regression model to evaluate PE effectiveness, by means of interviewing teachers 3 school districts in the United States. In this model, school facility resources were rated into 3 levels, according to teachers' satisfaction on 6 facility types. In essence, the existing researches have involved the assessment or classification on the circumstances of SSF by the quantities and subjective satisfaction. However, few writers have been able to draw on any systematic research into the size factor of SSF. Meanwhile, extensive researches have been carried out in both developed and developing districts, but the issue in mainland China has been rarely documented.

Shanghai, located in the east China, is one of world financial, scientific and innovation centers. The primary education resources and school infrastructure in Shanghai manifests conspicuous regional characteristics in suburban and urban districts, owing to the imbalance of population and economic development. Thus, SSF in Shanghai could be deemed as the epitome of world. The major objective of this study was to utilize a novel research method based on machine learning, for the investigation and evaluation of the SSF condition in primary schools (PSSFC) according to otherness in terms of the school type and location.

2. Materials and Methods

2.1. Subjects

All investigated samples were recruited from Shanghai, including 845 primary schools in 7 urban districts, 8 suburban districts and 1 urban-suburban fringe district. Stipulated by the Law of the People's Republic of China on Education, the subjects were classified into three school types: primary school (5-year curriculum in Shanghai, PS, N = 666), nine-year schools (Grade 1 to 9 combined, 9CS, N = 150） and twelve-year schools (Grade 1 to 12 combined, 12CS, N = 29).

2.2. Data Collection

The structured data analyzed in this research were imported via the pecialized data interface from the Sixth National Sports Facility Census (NSFC) database. Conducted by the General Administration of Sport of China in 2014, NSFC was intended to survey the quantity, distribution and purpose of the existing sports facilities in mainland China, and promote the coordinated development of sports and social economy. All kinds of sports facilities in various industries and forms of ownership were enrolled in with certainty. The consequent data processed by tabulation and collation were uploaded to the NSFC database, and reported to the local and national authorities.

In addition, the geographical data of SSF in NSFC was expressed as the address in the general information and location in the sports facility information. The latter referred to the surrounding of a particular facility. In view of the universe of this paper, the geographical character of each SSF cannot be directly manifested by both of the above indicators. Furthermore, the Pudong District, the second largest and unique urban-suburban fringe district in Shanghai, is allocated the most fundamental education resources (4 times than the largest Chongming District [27]). As a consequence, to promote the effectiveness of research, the geographic location (the longitude and latitude coordinate) of the measured schools are crawled by programing from the Baidu Map.

2.3. School Facility Condition Measurements

The statistical indicators of NSFC covered the all aspects of sports facility and were divided into 3 categories: general information (ownership type, ownership details, address and athletic team service information), sports facility information (type, scale, number of staff, location type, year of built and open service) and operation information (operation mode, hosted sports events, fitness training programs, operation income and costs). Owing to the characteristics of primary education, the operation information was omitted in this study whereas the rest relating to this paper were elected and categorized. The details of each descriptive measurement are described as follow:

2.3.1. Variety

Variety was assessed by the aggregated numbers of sports facility types in each school. In the statistic caliber of NSFC, the types of SSF were investigated listed in Table 1. The numbers of total facility types (FN), indoor facility types (IN) and outdoor facility types (ON) were calculated separately, in order to describe the condition of SSF precisely.

Table 1. Classification and Types of SSF.

Category	Types	Total Types
Outdoor	stadium, track field, soccer pitch (futsal and 7-a-side pitch included), basketball court (3-a-side pitch included), volleyball court, badminton court, tennis court, swimming pool, American football/rugby pitch, hockey pitch, table tennis table, handball field, cricket pitch, gate ball court, baseball field, archery field, bocce courts, skating field, skateboarding/roller skating field and fitness equipment	20
Indoor	track field, fitness room, basketball court, volleyball court, handball field, gymnastics room, badminton court, table tennis room, martial arts room, fight event training room, fitness room, yoga room, weightlifting room, fencing room, chess and card room, bowling room, futsal pitch, tennis court, hockey pitch, archery field, equestrian field, ice hockey field, skating field, curling kettle field, skateboarding/roller skating field, squash room and gate ball room	27

2.3.2. Size

The size of SSF was determined based on the area of individual facility. The area was defined as the effective area for training, competition and fitness activities, including the safety, buffer and accessibility area. To demonstrate the variance of SSF, the area of indoor (IA), outdoor (OA) and total facility (FA) and those average values were investigated.

2.3.3. Location

School location (SL) was presented as the distance of individual school to the Shanghai Municipal Peoples' Government, computed by using the Haversine formula [28]. Thereby, SL of all samples were divided to the urban (the distance less than equivalent to 15 km, $N = 385$) and suburban (the distance greater tha 15km, $N = 460$) in accordance with the General Development Plan of Shanghai City announced in 2014.

2.4. Data processing and Analysis

The general framework of analysis in this paper is illustrated in Figure 1. All SSF data were processed and modelled using Data Analysis Library (Pandas) for Python (Python 3.7.4 version for Windows, Python Software Foundation, Delaware, USA). The location of samples were gathered, and consolidated with the PSSFC as a 7-dimensional SSF dataset. After the standardization by the Min-Max normalization, the constructed dataset was projected to a lower two-dimensional space by using t-stochastic neighbor embedding (t-SNE) [29][30], which is a branch of the Manifold Learning. Lastly, we introduced an unsupervised machine learning method, Fuzzy C-mean (FCM), to cluster primary schools by the similarity in PSSFC and sort to five levels.

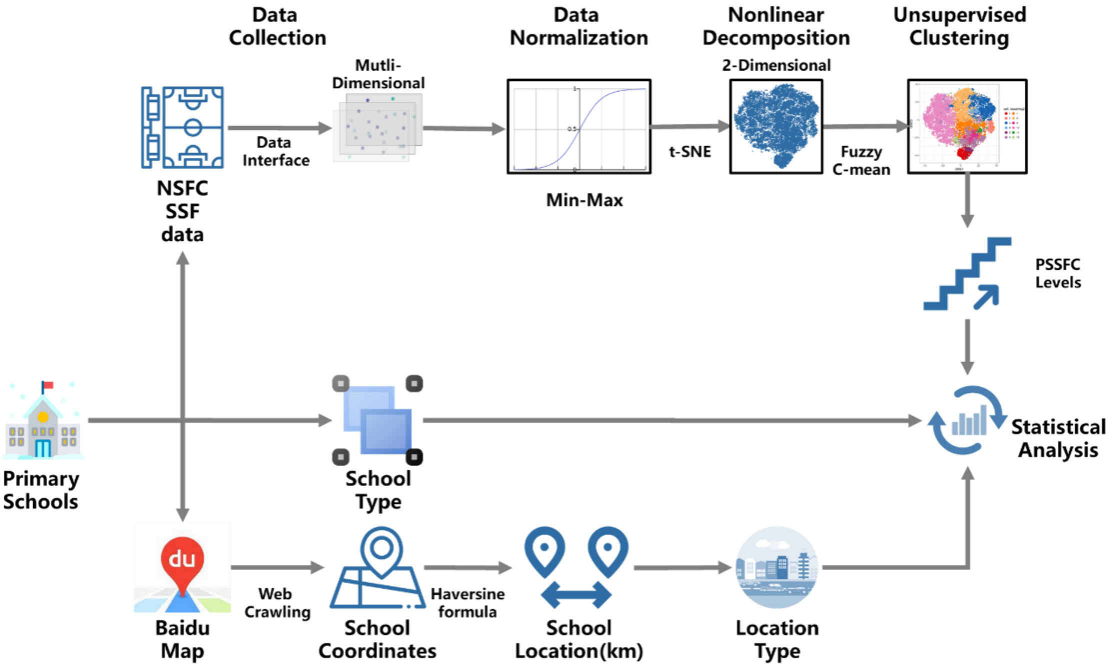


Figure 1. Overview of data processing and modeling pipeline. The raw data from NSFC and Baidu Map via internet were formatted as XML files, aligned by the school name converted into constructed Comma-Separated Values format. The combined dataset was rescaled on account of the different orders of magnitude in SSF square footage and SSF variety.

2.4.1. Nonlinear Unsupervised Decomposition

The original dataset was in high dimension with the complexity of visually feature recognition, from which 4 numeric segments featuring the PSSFC were extracted: overall (FN and FA), indoor (IN and IA), outdoor (IN and IA) and SL. In common sense, the overall segment was the summation of the indoor and outdoor, which shows linear relation. Inversely, the relation between indoor and outdoor segments exhibited nonlinear and independent. As a result, linear dimensionality reduction techniques (Principal Component Analysis, Linear Discriminant Analysis, Singular Value

Decomposition and et al) incapably displays the clarified representation of sample schools when reducing the high dimension down to 2.

As an alternative to this issue, a nonlinear manifold learning algorithm, t-SNE, was applied to reduce the dimension of dataset and visualize PSSFC in the lower dimension in favor of the successive evaluation. Innovated by the Stochastic Neighbor Embedding (SNE) [31], t-SNE uses a simpler-gradient-based symmetrized SNE cost function and replaces the Gaussian distribution with Student-t (a heavy-tailed distribution) in the computation in the lower-dimensional space, improving the crowd problem and the optimization problem.

The ideal of t-SNE, an unsupervised machine method, is transforming the similarity of the reduced sample points into the possibility: the Gaussian distribution in the original space (the higher dimensional space) and the Student-t distribution in the embedded space (the lower space). More specifically[30], given the N-dimensional SSF dataset *X* with n samples(x_1, x_2, ..., x_n), for each x_i chosen, define the conditional probability $p_{j|i}$ of picking another datapoint x_j in the samples as a neighbor to be proportional to the probability density of a Gaussian centere at x_i

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)} \tag{1}$$

in which, σ_i^2 is the variance of the Gaussian center on the datapoint x_j. In addition, $p_{i|i}$ is set to zero, because of concerning the similarity of two non-repetitive datapoints. The projection of x_i and x_j into the lower dimension expressed as y_i and y_j accordingly. For all datapoints in X, it is possible to find $Y^{(T)} = \left\{y_1, y_2, \dots, y_n\right\}$ in the lower dimensional space, which is the two-dimensional representation of *X*. The conditional probability of y_i and y_j, defined as q_{ij}, can be computed by

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right)} \tag{2}$$

Likewise, $q_{i|i}$ equals to zero. If the effect of decomposition is acceptable and the local features are preserved informatively, $p_{i|j}$ will be equivalent to $q_{i|j}$. Thus, the Kullback-Leibler divergence (K-LD)[32] is used to measuring the mismatch between $p_{i|j}$ and $q_{i|j}$. In t-SNE, the Student-t distribution (joint probability distribution) is applied rather than Gaussian (conditional probability distribution). Alternatively, t-SNE minimized K-LD between joint probability distribution, *P*, in the high-dimensional space and a joint probability distribution, *Q*, in the low-dimensional space:

$$\text{Min } KL\left(P\|Q\right) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3}$$

where the pairwise similarities in the low-dimensional and high-dimensional space are

$$q_{ij}=\frac{\exp\left(-\left\|y_i-y_j\right\|^2\right)}{\sum_{k\neq l}\exp\left(-\left\|y_k-y_l\right\|^2\right)} \tag{4}$$

$$p_{ij}=\frac{\exp\left(-\left\|x_i-x_j\right\|^2/2\sigma_i^2\right)}{\sum_{k\neq l}\exp\left(-\left\|x_k-x_l\right\|^2/2\sigma_i^2\right)} \tag{5}$$

This is referred as the symmetric SNE, the gradient of which is given by

$$\frac{\delta C}{\delta y_i}=4\sum_j (p_{ij}-q_{ij})(y_i-y_j) \tag{6}$$

The algorithm of t-SNE can be conducted as Table 2. $Y^{(t)}$ presents the solution after the t times iteration ,and $\alpha(t)$ is the momentum of t times iteration. When it comes to application, there exit serval parameters to be configured: the perplexity (*Perp*) of t-SNE is the cost function parameter indicating the numbers of effective neighbors for a particular datapoint and is given to discover proper σ by using binary search; The learning rate (η) denotes the velocity of the algorithm,

For the issue of SSF data, we set the configuration of previous parameters with $Perp = 25$, $\eta = 300$, $t = 800$. It should be clarified that in spite of the high complexity scaling ($O(n^2)$) compared with PCA and LDA, t-SNE performs with excellent performance in the visual description of SSF features in primary schools, on account of the finite numbers of measurements and samples. Moreover, the K-LD was applied to validate the effectiveness of t-SNE, the less value of which implied the less losses of original information after the embedding.

Table 2. The algorithm of t-SNE[30]

Algorithm of t-SNE

```
begin
    compute pairwise affinities $p_{j|i}$ with perplexity $Perp$
    set $p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}$
    sample initial solution $Y^{(0)}=\{y_1,y_2,\dots,y_n\}$
    for $t$ =1 to $T$ do
        compute low affinities $q_{ij}$
        compute gradient $\frac{\delta C}{\delta Y}$
        set $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
    end
end
```

2.4.2. Unsupervised Clustering

Clustering is to distinguish the objects with the similar attributes by allocating into different groups (clusters) [33]. In this paper, an unsupervised machine learning method, FCM algorithm [34], was introduced to dispose the discrimination of PSSFC. Unlike the traditional K-mean algorithm, FCM based on the fuzzy theory optimizes the objective function to obtain the membership degree of each sample point for all class centers and determine the category of sample points. For the realization on this issue, given the embedded SSF dataset *X*= {*x*1, *x*2, ..., *x*n} (in 2 dimension) and *j* clustering centers *C*= {*c*1, *c*2, ..., *cj*}, a *2* ×*n* matrix *U* can be generated to describe the clustered result：

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1j} \\ u_{21} & \cdots & u_{2j} \end{pmatrix} \tag{7}$$

where, u_{ij} ranged from zero to one is the degree of membership for individual xi belonging to category *j*, which can be calculated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{8}$$

where, *m* is a real number greater than 1. After the *t* times iteration FCM attempts to minimize the cost function as below:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m \left\| x_i - c_j \right\|^2, 1 \le m < \infty \tag{9}$$

The FCM terminates until the following condition is compromised:

$$\max_{ij} \left\{ \left| u_{ij}^{(t+1)} - u_{ij}^{(t)} \right| \right\} < \varepsilon \tag{10}$$

which implies that the result has achieved comparative (local or global) optimum, the degree of membership will not alter significantly if the iteration continues.

Generally, the numbers of cluster center *c* in FCM is a predetermined parameter [35]. However, lack of standards or guidance in the issue of PSSFC evaluation has been existed for the current researches. Therefore, the numbers of cluster center were adjusted from one to ten with the purpose of finding the best classification of SSF. The fuzzy partition coefficient (FPC)[36] was selected as the validity index of clustering shown in Equation 11. The others parameters in FCM were *m* = 1.5, *t* = 1000, ε = 0 .001.

$$FPC=\frac{\frac{1}{c}\sum_{i=1}^{N}\sum_{j=1}^{c}u_{ij}^{2}-1/c}{1-1/c} \tag{10}$$

2.4.3. Statistical Analysis

The original data were labelled in a tiered scale after the process of t-SNE and FCM. It was specified that the individual with Level-A indicated the excellence of PSSFC in such primary school. The description of all tiers was exhibited with descriptive statistics (means and standard deviations, mean±SDs). Furthermore, a two-way analysis of covariance, controlling for the type and location of primary schools, was performed on to figure out if there exited differences on the variety and size in each PSSFC level, with the significant level at 0.05. All analyses were processed using Statistical Package for the Social Sciences (SPSS) 22.0 on Windows (IBM, Chicago, IL, USA).

3. Results

3.1. Primary Education Resource in Shanghai, China

Most primary schools (78.9%) in Shanghai were attributed to the five-year primary education system, and the nine-year system schools (9CS) occupied over one-sixth of the total. As to the geographical distribution pattern, the urban primary schools and the suburban were approximately equal in quantity with the ratios of 1:1.07.

Table 1. Overview of Primary Education in Shanghai, China.

Primary School Type	Urban	Suburban	Total
12CS	13	16	29
9CS	89	61	150
PS	306	360	666
Total	408	437	845

3.2 Visualization and Gradation of PSSFC

3.2.1. Visualization of PSSFC in 2D space

Figure 2 shows the embedding of PSSFC into two-dimensional space, the space (y_1, y_2), for all of the 845 sampled primary schools. The data with dissimilarity are isolated in the embedded space, while the samples of the similar attributes can be almost assembled into the adjacent area. This hints that the visualization of PSSFC using t-SNE is capable of observing the global overview of PSSFC.

In our experiment, K-DL equals to 0.3032, which denotes that the dimensional reduction via t-SNE basically preserved the features of original PSSFC dataset.

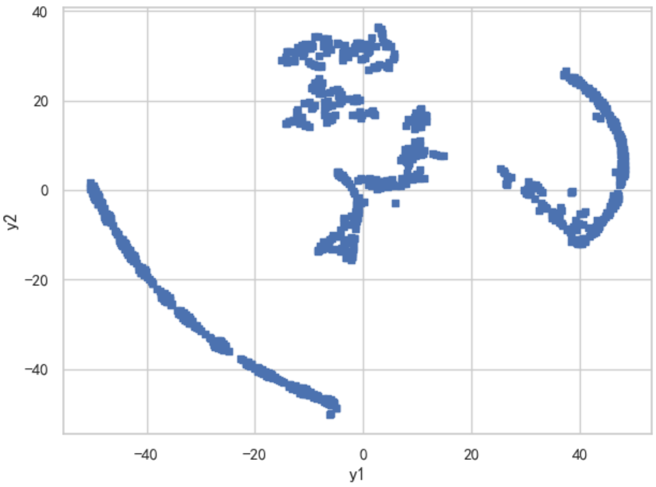
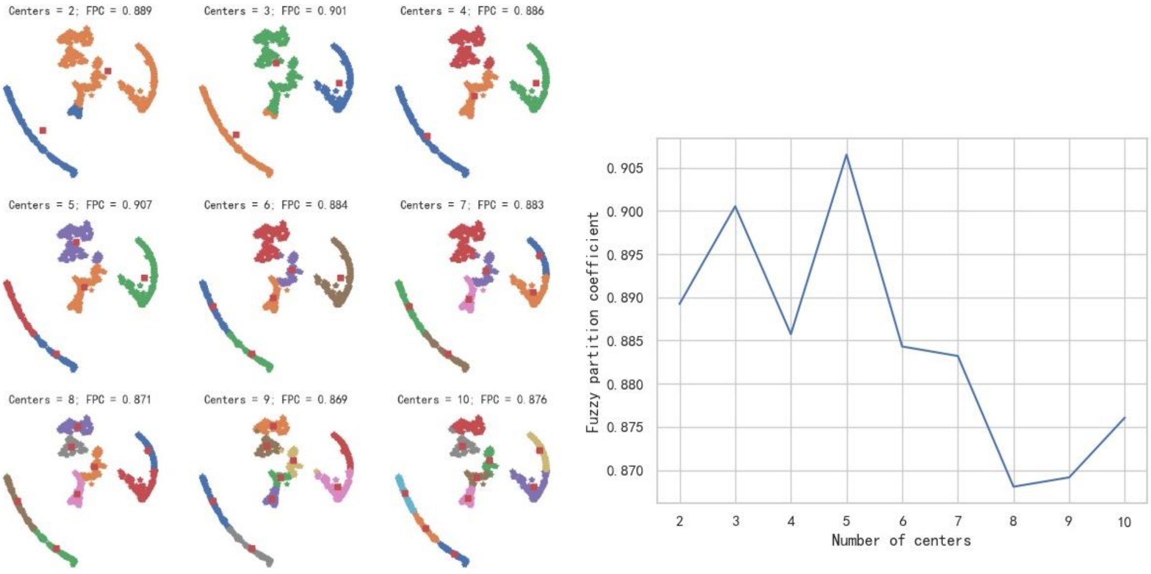


Figure 2. Embedding of PSSFC into two dimensions via t-SNE.

3.2.2. Result of Unsupervised Clustering



(**b**)

(**a**)

Figure 3. Clustering of PSSFC via FCM. (**a**) Color coding on the results of FCM visually exhibits the effect on occasion of different numbers of clustering centers ranged from one to ten. (**b**) The comparison of FPC for selection of number of centers *c*

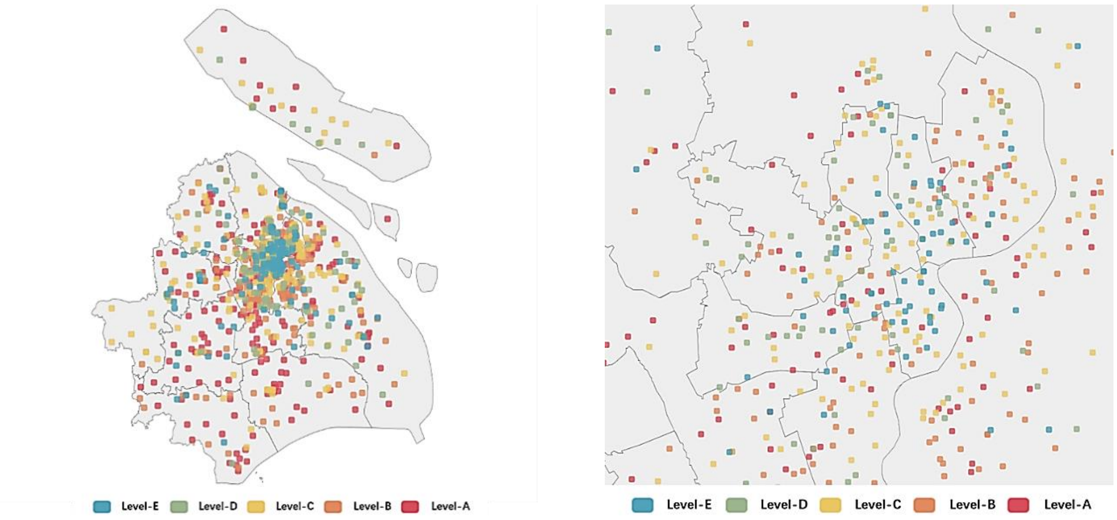
Figure 3 illustrates the results of FCM clustering on the embedded PSSFC via t-SNE. PSSFC in 2D space is intuitively supposed to be classified to three categories at the first glance from Figure 2. Nonetheless, It can be recognized from Figure 3(a) that some datapoints, located near the periphery of each category, are inaccurately divided under the situation of three or four centers chosen. More quantitatively, in this study, the validation of clustering was conducted via computing the FPC objectively. It can be concluded from Figure 3(b) that FPC is ramped when c is chosen from 1 to 4. The performance of FCM reaches the optimum (0.907) with *c* = 5. Afterwards, the effect of FCM clustering declines dramatically.

3.4 Location and School Type characteristics of PSSFC

Table 3. Descriptive Statistics on Five-levelled PSSFC in Using Machine Learning

Location	School Type	PSSFC Level										Total
		Level-A		Level-B		Level-C		Level-D		Level-E		
		N	%	*N*	%	*N*	%	*N*	%	*N*	%	
Suburban	12CS	9	69.2%	3	23.1%	1	7.7%	0	0.0%	0	0.0%	13
	9CS	56	62.9%	12	13.5%	16	18.0%	1	1.1%	4	4.5%	89
	PS	78	25.5%	51	16.7%	91	29.7%	45	14.7%	41	13.4%	306
	%	35.1%		16.2%		26.5%		11.2%		11.0%		408
Urban	12CS	9	56.3%	3	18.8%	2	12.5%	1	6.2%	1	6.2%	16
	9CS	21	34.3%	14	23.0%	13	21.3%	4	6.6%	9	14.8%	61
	PS	43	11.9%	86	23.9%	101	28.1%	70	19.4%	60	16.7%	360
	%	16.7%		23.6%		26.5%		17.2%		16.0%		437

Notes: 12CS = 12-year school; 9CS = 9-year school; PS = 5-year school, primary school only.



(**a**)

(**b**)

Figure 4. PSSFC in different district in Shanghai labelled with gradients, among which red is Level-A, green is Level-E). (**a**) The geographical distribution panorama of primary schools and its PSSFC in Shanghai. (**b**) The PSSFC in urban districts zoomed in to the scope of urban area (distance ≤15 kilometers to the Shanghai Municipal Peoples' Government).

Observing PSSFC in the scope of geographical distribution, we present the colored-separated geographic scatter diagram to illuminate the impact of location on PSSFC. In Figure 4, the PSSFC Level of primary schools is marked by the gradients ranged from green to red. To further support the findings, Table 3 is listed to compare the calculated statistical quantity and proportion in the respect of spatial and phyletic distribution. Combining the two results, over half of primary schools in the suburbs are deployed with upper-class PSSFC (above Level-C), the proportion of which is strikingly higher than in the inner city by around 11%. As a result, PSSFC in urban schools seems to be a bit insufficient to compromise the demand of children's sport participation, especially for five-year school (aggregated 36.1%). It can be summarized that PSSFC in suburban districts generally is better, although percentage of Level-C between urban and suburban is equal. Another striking result to emerge from the Table 3 is that PSSFC of PS remains inferiority to t 9CS and 12 as a whole.

3.4 Detailed Description on PSSFC in Different Level

According to the pervious clustering results, the description of PSSFC rated from Level-A to Level-E (shown in Table 3) was generated by the projection to the original dataset for the purpose of improving the interpretability of machine learning. Approximate a quarter of all the subjects are labelled with Level-A (*N* = 218, 25.56%) and Level-C (26.51%) on PSSFC. In contrast, there are no difference on the percentages between Level-D and Level-E.

It is obvious from the table that the stepwise reduction on PSSFC from Level-A is the dominant properties of the evaluation. Primary schools in the Level-A tier with at least 1 indoor sports facility, possesses the most types (4.58 in total, 3.25 in outdoor and largest area (over 2 times than the Level-B) of sports facility. The variety and size of the sports facility in Level-B seems to be the nearly one-third to half diminishment of the Level-A. The similar result can be figured out between the Level-B and Level-C. Interestingly, the Level-C and Level-D slightly exceed on average total and outdoor area than the upper class. Ultimately, the majority of measurements between Level-D and Level-E perform a high similarity, with the exception on total and indoor area of sports facility.

Table 4. Descriptive Statistics on PSSFC in Five Levels Using Machine Learning

	Level-A	**Level-B**	**Level-C**	**Level-D**	**Level-E**
Schools (*N*)	218	168	224	121	114
Percentage (%)	25.56%	20.00%	26.51%	14.32%	13.61%
Total Facilities					
Types (*N*)	4.58±1.34$^{+\times}$	2.99±0.23$^{+\times}$	1.99±0.09	1±0	1±0
Area (m^2)	9036.3±5853.56$^{*+\times}$	4184.93±2034.31^{*+}	3630.91±2261.84$^{*+\times}$	2984.8±1338.26$^{*+\times}$	851.58±357.23$^{*+\times}$
Avg. Are a(m^2)	2056.66±1333.1^{*+}	1401.51±677.68^{*+}	1839.44±1233.34$^{*+\times}$	2984.8±1338.26$^{*+\times}$	851.58±357.23$^{*+\times}$
Indoor Facilities					
Types (*N*)	1.33±0.93^{+}	0.82±0.55*	0.33±0.47*	0.02±0.13	0±0
Area (m^2)	765.03±715.68$^{*+\times}$	339.97±362.66$^{\times}$	135.47±252.08$^{*+\times}$	3.21±25.62	0±0
Avg. Area (m^2)	534.19±398.87^{+}	315.77±342.21$^{\times}$	135.47±252.08$^{*+\times}$	3.21±25.62	0±0
Outdoor Facilities					
Types (*N*)	3.25±1.23$^{*\times}$	2.18±0.6*	1.66±0.494*	1±0	0.98±0.13$^{*+\times}$
Area (m^2)	8271.27±5601.39$^{*+\times}$	3844.97±2032.21^{*+}	3495.44±2236.85$^{*+\times}$	2984.8±1338.26$^{*+\times}$	848.37±363.95^{*+}
Avg. Area (m^2)	2747.19±1815.75^{*+}	1802.12±912.82^{+}	2234.83±1628.06^{*+}	2984.8±1338.26$^{*+\times}$	848.37±363.95^{*+}

Notes: Avg. Area = average area of facilities; SD = standard deviation; Data are expressed as mean ± SD; * Significant difference between urban and suburban primary schools with the same PSSFC Level ($p < 0.05$); + Significant difference among different school types (PS, 9CS, 12CS) with the same PSSFC Level ($p < 0.05$); × Significant interaction of the types and location of schools between the PSSFC variables with the same PSSFC Level ($p < 0.05$) .

The relation between the location of school and measured PSSFC variables was tested, given the equivalent PSSFC level and type of schools. In Table 4, the location of primary school is decisive to its outdoor facility, variety of indoor facility in Level-B and Level-C, and total size in all level ($p < 0.05$). There exists no significant association between the location and indoor facility below Level-D ($p > 0.05$). Likewise, the overwhelming majority of PSSFC variables above Level-C (except indoor facility in Level-B) and the outdoor below the intermediate level have significant associations with the individual's type. This indicates the school educating more grades is deployed with better sports environment ($p < 0.05$). Considering the action of combined location and school type, these two factors have impressive impact on the most variables in Level-A, overall size in all below Level-B, area of indoor facility above Level-D ($p < 0.05$).

4. Discussion

At present, vast studies merely investigated the correlation between the children's PA and PSSFC, rather than focusing on the PSSFC by multi-dimension analysis. The initial purpose of our research is to utilize the regional representative data from the sport environment of primary schools in the most rapidly developing country, generating an interpretable and symbolic PSSFC classification criteria via the machine learning. Thereby, this study explores a tentative methodology in the domain of school sports environment. Although some supervised linear modeling method (the existence of on-site facilities as variables of evaluating school environment) [38] was applied into this issue, the absence of systematic and precise classification is deemed to be solved. Under the situation of missing the authorized guideline, the evaluation on PSSFC can be attributed to the unsupervised learning problem, according to the classification of machine learning [39]. In consideration of the non-linear relationship between indoor and outdoor PSSFC factors, t-SNE was taken into account to settle the high-dimensional and non-linear PSSFC dataset, on the basis of its superior performance [40-43]. Our finding confirmed that the unsupervised machine is capable of retaining the original features and presenting the PSSFC characteristic in more visual ways. To obtain the aim of evaluation on PSSFC, FCM performing more flexibility and robustness compared with K-means, was verified in this paper, and its result of clustering into 5 gradient levels comprehensively showed the interclass differences of PSSFC.

Prior studies have noted that the adequate in both diversity and space of PSSFC is crucial to children's PA. Ridgers *et al.* [44] pointed out that the space of children's' PA had positive correlation with both the duration and intensity of physical activity. The analogous findings were proved in Japan, which is similar to China: children in the high-facility primary schools participated more MVPA than those in the low-facility group [22]. In other words, regardless of gender, lack of sports facilities is a vital barrier to the participation of children's PA [45-47]. According to the gender disparities of the preference on sports facilities, it is recommend that besides the space of PSSFC, the variety ought to be considered as well. More precisely, boys dominate the outdoor sports facilities, such as soccer pitch, with girls preferring the indoor activities such as dancing and gymnastics [48]. As seen in Table 4, the quantified evaluation on PSSFC proposed was consistent with the above finding, which suggests that a particular primary school with PA-demand-compliant PSSFC should be configured with 2 types of sports facilities (including 1 indoor is preferred) and at least 3000 m^2 accessible area. This on aspect identifies the effectiveness and validity of the two-approach-combined unsupervised machine learning.

Regarding the impact of urbanization and school type, the result of this study at first implies the PSSFC has a striking association with the individual location, especially for the aggregated sports space and outdoor condition. The similarity was proved in Taiwan [49] and Ontario, Canada [51]. Although the urban PSSFC probably affected by the shortage of land resources in the developed city seemed to be inferior to the suburban, the high density of inner-city sports facilities, particularly the indoor facilities, supplement to the accessibility of children's PA [51-54]. To the best of our knowledge school-type differences have not been assessed in other studies. Out study proved the significant links between PSSFC variables and individual school type. In the usual sense, among all three types, 12CS catering 7 to 12 years old student is the most populous type, requiring the best PSSFC for multigrade-shared PA in the recess and afterschool. A possible explanation for these results may be the relevant regulation of primary education in mainland China [55]. It stipulates that a PA room, greater than 40 m^2, must be deployed in the urban 9CS and 12CS, which is not the essential requirement of PS. This explains the association between indoor facilities area and school types.

Several limitations of this study should be clarified. First, the analysis is a representative cross-sectional in the developed region, which is incapable of providing a comprehensive review of PSSFC in mainland China. Second, the approach of visualization and evaluation on PSSFC is exclusively based on the objective data. Due to the vacancy of indicators in the NSFC, the following potential factors should be involved in further practical application and studies: the subjective assessment, the student population in each school, PA space per capita the land occupation and floor space of school. Despite these limitations, the current study provides a demonstration of interdisciplinary research between artificial intelligence and PA environment, and analyzes the spatial and phyletic characteristics of PSSFC in the developing country. The evaluation method on PSSFC can be used to inform the regionalized and personalized intervention for children's PA promotion, according to the current state of individual PSSFC.

5. Conclusions

The finding in this paper verified that the combined approach unsupervised learning facilitates to understanding the characteristics of sport facility in primary schools in a visual way. The approach used in this study with acceptable levels of reliability and interpretability and was found to differentiate sport facility in primary schools effectively. In addition, with the statistical analysis on the classification results, the spatial and typological association with the variables of sports facility in primary schools was discovered and has the consistency with the previous researches. Our findings revealed that the regional school sport facility condition in primary schools has the strict geographic distribution characteristics. A primary school with more grades of students is configured with more types and sizes of sports facility

Further work is needed to quantify its association with the subjective remarks, attending numbers of students and other potential influences and in more varied settings.

Author Contributions: Jun Xia participated in the design, conducted the programming, modelling and statistical analysis and drafted the manuscript; Pei-Jie Chen administrated and supervised the study, and reviewed the manuscript; Ji-Hong Wang assisted in gathering the data; Zhen-Bo Cao and Jie Zhuang helped in conducting the study and revise the manuscript; Qiang Zhang helped to manage and analyze the data. All authors read and approved the final manuscript.

Acknowledgments: This study is supported by the Grand-in-Aid for Postdoctoral Scientific Research from China Postdoctoral Science Foundation (No. 2018M632148) and the Research Projects from Shanghai Shanghai Association for Science and Technology (No. 19080503000), and Decision Consulting Project of from the Development Research Center of Shanghai Municipal People's Government (No. 2018-Z-55).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kelly, I.R..; Phillips, M.A.; Revels M.. Contribution of the School Environment to Physical Fitness in Children and Youth. *J. Phy.s Act. Health* **2010**, *7*, 333-342.
2. Troiano R.P.; Flegal K.M. Overweight prevalence among youth in the United States: why so many different numbers? *Int. J. Obes. Relat. Metab. Disord.* **1999**, 23, S22–S27.
3. Ogden C.L.; Flegal K.M.; Carroll M.D. Prevalence and trends in overweight among U.S. children and adolescents, 1999–2000. *JAMA* **2002**, 88, 1728–1732.
4. Gordon-Larsen P.L.; Adair S.; Nelson M.C. Five-year obesity incidence in the transition period between adolescence and adulthood: the National Longitudinal Study of Adolescent Health. *Am. J. Clin. Nutr.* **2004**, 80, 569–575.
5. Slyper A. The pediatric obesity epidemic: causes and controversies. *J. Clin. Endocrinol. Metab.* **2004**, 89, 2540–2547.
6. Biddle, S.J.; Gorely, T.; Stensel, D.J. Health-enhancing physical activity and sedentary behavior in children and adolescents. *J. Sports Sci.* **2004**, 22, 679–701.
7. World Health Organization. Global recommendations on physical activity for health. Geneva: World Health Organization; **2010**.
8. Ainsworth B.E. How physically active are our children? A global view. *J. Sport Health Sci* **2016**, 5, 400-1.
9. Lu C.; Stolk R.P.; Sauer P.J. Factors of physical activity among Chinese children and adolescents: a systematic review. *Int. J. Behav. Nutr. Phys. Act.* **2017**, 14, 36.
10. Rahman T.; Cushing R.A.; Jackson R.J. Contributions of built environment to childhood obesity. *Mt. Sinai. J. Med.* **2011**, 78, 49–57.
11. Fan X, Cao Z.B. Physical activity among Chinese school-aged children: national prevalence estimates from the 2016 Physical Activity and Fitness in China-The Youth Study. *J. Sport. Health Sci.* **2017**, 6, 388–394.
12. Liu Y.; Tang Y.; Cao Z.B. Results from the 2016 Shanghai's (China) report card on physical activity for children and youth. *J. Phys. Act. Health* **2016**, 13(Suppl. 2), S124–8.
13. Duan J, Hu H, Wang G, Arao T. Study on current levels of physical activity and sedentary behavior among middle school students in Beijing, China. *PLoS One* **2015**, 10, 1371–1383.
14. Bauman, A.E.; Reis, R.S.; Sallis, J.F. Correlates of physical activity: Why are some people physically active and others not? *Lancet* **2012**, 380, 258–271.
15. Robertson-Wilson J.E.; Leatherdale S.T.; Wong S.L. Social-ecological correlates of active commuting to school among high school students. *J. Adolesc. Health* **2008**, 42, (5), 486–95.
16. O'Malley P.M.; Johnston L.D.; Delva J.; Terry-McElrath Y.M. School physical activity environment related to student obesity and activity: a national study of schools and students. *J. Adolesc. Health* **2009**, 45(3Suppl), S71–81.
17. McGrath L.J.; Hopkins W.G., Hinckson E.A. Associations of Objectively Measured Built-Environment Attributes with Youth Moderate-Vigorous Physical Activity: A Systematic Review and Meta-Analysis. *Sports Med.* **2015**, 45(6), 841-865.
18. Jago, R.; Baranowski, T. Non-curricular approaches for increasing physical activity in youth: a review. *Prev. Med.* **2004**, 39, 157–163.
19. Haug E.; Torsheim T.; Sallis J.F. The characteristics of the outdoor school environment associated with physical activity. *Health Educ. Res.* **2010**, 25(2), 248–256.
20. Haug E.; Torsheim T.; Samdal O. Physical environmental characteristics and individual interests as correlates of physical activity in Norwegian secondary schools: the health behavior in school-aged children study. *Int. J. of Phy. Act. Behav. Nutr.* **2008**, 5(1), 47–56.
21. Haug E.; Torsheim T.; Samdal O. Local school policies increase physical activity in Norwegian secondary schools. *Health Promot. Int.* **2009**, 25(1), 63–72.
22. Ishii K.; Shibata A.; Sato M.; Oka K. Recess Physical activity and perceived school environment among elementary school children. *Int. J. Environ. Res. Public Health* **2014**, 11, 7195-7206.
23. Rezende L.F.; Azeredo C.M.; Silva K.S; Claro R.M. The role of school environment in physical activity among brazilian adolescents. *PloS One* **2015**, 10(6), e0131342.
24. Lo K.Y; Wu M.C.; Tung S.C. Association of school environment and after-school physical activity with health-related physical fitness among junior high school students in Taiwan. *Int. J. Environ. Res. Public Health* **2017**, 14(1), E83.
25. Huag E.; Torsheim T.; Sallis J.F. Title of Thesis. The characteristics of the outdoor school environment associated with physical activity. *Health Educ. Res.* **2010**, 25(2), 248-256.

26. Bevans B.K.; Fitzpatrick L.A.; Sanchez B. M. Physical education resources, class management, and student physical activity levels a structure-process-outcome approach to evaluating physical. *J. School Health* **2010**, 80(12):573-580.
27. The Statistics Bureau of Shanghai. *Shanghai statistical yearbook 2018*; China Statistic Press: Beijing, China, **2019**, pp. 154.
28. Chopde N.R.; Nichat M.K. Landmark based shortest path detection by using A* and Haversine formula. *Int. J. Innovative Res. Comput. Commun. Eng.* **2013**, 1(2), 298-302.
29. van der Maaten L.; Hinton G.; Visualizing high-dimensional data using t-SNE. *J. Mach. Learn.* **2008**, 9, 2579-2605.
30. van der Maaten L. Accelerating t-SNE Using tree-based algorithms. *J. Mach. Learn.* **2014**, 15, 3221–45.
31. Hinton G.E.; Roweis S.T. Stochastic neighbor embedding. Proceedings of Advances in Neural Information Processing, Systems; Cambridge, UK, 2002; MIT Press: Massachusetts, USA, 2002, pp. 833-840.
32. Kullback S.; Leibler R.A. On information and sufficiency. *Annals of Mathematical Statistics* **1951**, 22(1), 79-86.
33. Nayak J.; Naik B.; Behera H.S. Fuzzy c-means (FCM) clustering algorithm: a decade review from 2000 to 2014. (2015) Proceedings of the International Conference on CIDM; Florida, USA, 2014; Springer, New Delhi, India, 2015, (2):133-149.
34. Bezdek J.C.; Ehrlich R.; Full W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci*. 1**984**, 191-203.
35. Izakian H.; Abraham A. Fuzzy c-means and fuzzy swarm for fuzzy clustering problem. *Expert Syst. Appl*. **2011**, 38(3), 1835-1838.
36. Wu K.; Yang M. A cluster validity index for fuzzy clustering. *Pattern Recogn. Lett*. **2005**, 26(9), 1275-1291.
37. Kirby J.; Levin K.A.; Inchley J. Associations between the school environment and adolescent girls' physical activity. *Health Educ. Res.* **2012**, 27(1), 101-114.
38. Kruiger J.F.; Rauber P.E.; Martins R.F. Graph layouts by t-SNE. *Comput. Graph Forum.* 2017, 36, (3), 283-294.
39. Murphy K.P. Machine learning: a probabilistic perspective, 4th ed.; MIT Press: Massachusetts, USA, **2013**; pp. 213.
40. Olivon F.; Elie N.; Grelier G. MetGem software for the generation of molecular networks based on the t-SNE algorithm. *Anal. Chem*. **2018**, 90, (23), 13900-13908.
41. Platzer A. Visualization of SNPs with t- SNE. *Plos One* **2013**, 8(2), e56883.
42. van Unen V.; Hollt T.; Pezzotti N. Visual analysis of mass cytometry data by hierarchical stochastic neighbor embedding reveals rare cell types. *Nat. Commun.* **2017**, 8, (1), 1-10.
43. Dimitriadis G.; Neto J.P; Kampff A.R. t-SNE visualization of large-scale neural recordings. *Nueral Comput.* **2018**, 30, (7), 1750-1774.
44. Ridgers N.D.; Fairclough S.; J; Stratton G. Variables associated with children's physical activity levels during recess: the A-Class project. *Int. J. Behav. Nutr. Phy.* **2010**, 7(1), 74-79.
45. Stanley R.M.; Boshoff K.; Dollman J. Voices in the playground: a qualitative exploration of the barriers and facilitators of lunchtime play. *J. Sci. Med. Sport*. **2012**, 15(1), 44-51.
46. Parrish A.; Yeatman H.; IversonD. Using interviews and peer pairs to better understand how school environments affect young children's playground physical activity levels: a qualitative study. *Health. Educ. Res.* **2012**, 27(2), 269-280.
47. Thompson J.L.; Davis S.M.; Gittelsohn J. Patterns of physical activity among American Indian children: an assessment of barriers and support. *J. Community Health*. **2001**, 26, 423–445.
48. Pawlowski C.S.; Tjørnhøj-Thomsen T.; Schipperijn J. Barriers for recess physical activity: a gender specific qualitative focus group exploration. *BMC Public Health* 2014, 14, 639-646.
49. Chen J.; Unnithan V.B.; Kennedy C. Correlates of physical fitness and activity in Taiwanese children. *Int. Nurs. Rev.* **2008**, 55(1), 81-89.
50. Hobin E.; Leatherdale S.T.; Manske S. Are environmental influences on physical activity distinct for urban, suburban, and rural schools? A multilevel study among secondary school students in Ontario, Canada. *J. School Health* **2013**, 83(5), 357-367.
51. Huang J. Shen G.Q. Zheng H.W. Is insufficient land supply the root cause of housing shortage? Empirical evidence from Hong Kong. *Habitat. Int.* **2015**: 538-546.
52. Wong B.Y.; Cerin E.; Ho S.Y. Adolescents' physical activity: competition between perceived neighborhood sport facilities and home media resources. *Int. J. Pediatr. Obes.* **2010**, 169–176.

53. Prins R.G.; Ball K. Timperio A. Associations between availability of facilities within three different neighborhood buffer sizes and objectively assessed physical activity in adolescents. *Health Place* **2011**, 17: 1228–1234.
54. Reimers A.K.; Wagner M.; Alvanides S. Proximity to sports facilities and sports participation for adolescents in Germany. *Plos One* **2014**, 9(3), e93059.
55. Standards for construction of school buildings for urban primary and secondary schools. National center for schooling development program. Available online: http://www.csdp.edu.cn/article/589.html (accessed on 10 Oct 2019)