Faculty of Natural and Mathematical Sciences
Department of Information

King's College London
Strand Campus, London,
United Kingdom

7CCSMUIP

Individual Project Submission 2019/20

Name:	Shengwei Zhou
Student Number:	19046602
Degree Programme:	Urban Informatics
Project Title:	British School and Adolescent Suicide Ideation
Supervisor:	Angus Roberts
Word Count:	6583

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

☑ I **agree** to the release of my project

□ I **do not** agree to the release of my project

Signature: Shengwei Zhou **Date:** August 21, 2020

Department of Information
King's College London
United Kingdom

7CCSMUIP Individual Project

British School and Adolescent Suicide Ideation

Name: **Shengwei Zhou**
Student Number: 19046602
Course: Urban Informatics

Supervisor: Angus Roberts

This dissertation is submitted for the degree of MSc in Urban Informatics.

Acknowledgements

First of all, I would like to thank my supervisor Angus for his supervision and care for me over the past six months. He has been giving me careful guidance from the beginning of the topic selection, helping me determine the research direction and promote the progress and development of my thesis. Secondly, I would like to express thankfulness to Mr. Xingyi Song from Sheffield, developed the C-LSTM-CNN model which was the foundation of my research. He also provided me with enthusiastic help and guidance, took time to give me video guidance during the epidemic, solved the dilemma I encountered in the code and algorithm. Thanks to CUSP and all the tutors in CUSP, especially Yijing and Claire, who have given me a lot of help both in school and life. Thanks to Shuangshuang Jin, Di Jing, Hanfeng Xiao, Yue Chen, Yuewei Jiang, Yulin Lai and other classmates and friends who have been with me in the front line when I was struggling hard.

This year has been a very difficult year. If there was no COVID-19, I should be still in my apartment in Farringdon, lying on my cot to finish my entire school year in the UK. COVID-19 has given me a completely different experience of studying abroad. Although I have only spent nine months in the UK, it has left me with a very deep memory. Thanks to COVID-19 for saving me time and money when traveling, which also made me missing the opportunity to appreciate the many beautiful scenery in the UK. Therefore, I also want to thank my parents, who accompanied me and kept silently supporting and encouraging me for the last six months when they were supposed to be happily enjoying the two-person world without me.

I also want to thank my friends Jie Chen, Xin Chen, Zhuoqing Chen, Peichen Liu, Zhuoyang Jiang and others. During this half year in China, whenever I feel upset, they will always listen to my complaints and entertain me, Let me regain the courage to fight. I also feel that my good friend Jiaming Lu who gave me help on the paper. Thanks to my best friend Yixuan Liu from the United States for crossing oceans with me in the past year, transnational video and phone calls in transnational time zones. Every time I chat with you, I can feel your vitality and happiness. Here I wish you a happy engagement. I wish you a happy marriage.

Finally, I would also like to thank my boyfriend Tianle Wang. Whether I were in UK or China, he has always been with me, accompany me to sleep with good morning, and accompany me at sunset with good night, letting me know that he will always be behind me even in the most difficult and sad period.

Abstract

Suicide, especially adolescent suicide, is often prone to clustering in geographic space, but there are few geographic spatial correlation analyses on suicidal consciousness and suicidal tendency. In this study, the sentence level-based C-LSTM-CNN algorithm was used to label the suicidal tendency for 23,455 patients from Child and Adolescent Mental Health Services (CAMHS), and the labeling results were combined with the data of the primary school in London to perform spatial autocorrelation and clustering analysis on geospatial level.

Nomenclature

NLP	Natural Language Processing
CAMHS	Child and Adolescent Mental Health Services
EHRs	Electronic Health Records
ASD	Autism Spectrum Disorder
LSOA	Lower Layer Super Output Area
LISA	Local Indicators of Spatial Association
SLaM	South London and Maudsley NHS Foundation Trust
CRIS	Clinical Record Interactive Search
IQR	Interquartile Range

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

According to the'Suicide Fact Sheet' published by WHO in 2019, suicide has become the third leading cause of death among young people aged 15-19, causing a death for every forty seconds [1]. Mental health problems are the main risk factors for attempted suicide. For example, among people who attempted (or completed) suicide in the United States in 2010, approximately 44% were diagnosed with mental health problems, while 31% were receiving mental health care.

Serious suicidal thoughts or completed suicide attempts may cause serious psychological effects, and even more empirical consequences, such as subsequent suicide attempts or even completion of suicide. In the general population, prior suicide attempt is the leading risk factor for suicide. Suicide is a serious public health problem; however, suicide can be prevented through timely, evidence-based, and often low-cost interventions. In order for the national response to be effective, a comprehensive multisectoral suicide prevention strategy needs to be developed.

Clinically, identifying young people with suicidal tendencies, attempted suicides, or self-harm is a typical method of personal prevention and suicide attempts. Self-harm is a key factor related to suicide among young people [2], so it is often regarded as a precursor to suicide. Furthermore, since suicide attempts are often linked to mental illnesses [3], before suicide attempts occur, the possible preventive method to prevent suicides is to identify, treat and care for people with mental and substance use disorders early. Therefore, it is very important to study the pathological changes of adolescents and the recognition of suicide awareness and expectations, and even subsequent suicide prevention and intervention.

A large number of studies have proved that suicides among teenagers are more likely to be clustering than adults [4, 5]. Especially in certain locations, including institutions (for example, schools, universities, psychiatric hospitals, and juvenile offenders), the number of suicides may be higher than expected. They may also involve related events that are geographically dispersed. The locations exposed in the cluster may be threatened by future clusters. The mechanisms involved in clusters include social communication (especially through inter-personal communication and media communication). It recognizes that suicidal behavior is widespread, and that susceptible young people are likely to interact with other suicidal people. Social cohesion promotes thought and The spread of attitude. The impact of suicide clusters on communities and institutions is usually far-reaching.

A paper published this year in The Lancet Child and Adolescent Health examines ways to establish clusters within communities and institutions and provide corresponding support and effective interventions [4]. By monitoring suicidal behavior, identifying susceptible groups, target group intervention and national intervention to identify suicide groups. But what if we can identify suicide groups earlier and even before the first suicide incident? In addition to the gathering of suicide incidents, is it possible that suicidal tendencies or suicidal consciousness may also occur between communities and schools?

1.2 Motivation

This study attempts to do a related geospatial analysis of suicidal tendencies in schools before suicide or suicide attempts occur to explore the clustering phenomenon behind them. Before the spatial analysis takes place, the identification of adolescent suicidal tendencies based on Natural Language Processing (NLP) and deep learning will be completed first to provide relevant data for subsequent steps. The focus of this article is not to explore the deep reasons behind youth suicide, nor to simply describe the geospatial mapping of youth suicide, but to analyze the geospatial relationship between youth suicide tendency and school, and to explore the linkage between the two.

The use of predicted suicidal tendencies instead of the existing suicide records data is that the suicidal tendencies predicted by the deep learning model can significantly determine whether the patient is at risk of suicide, thus providing support for suicide prevention and intervention to a certain extent. On the other hand, by analyzing the data of the prediction of suicidal tendency at the geospatial level, we can better judge the multiple relationship between the suicidal tendency in geospatial and the school, so as to better provide support for suicide prevention at the school level.

1.3 Report Structure

The report will begin with the background about both suicide clustering and the suicide iteation identification as well as related literature review. After introducing all previous work, the data and methodology used in this research will be introduced. In the next two chapters, the results of various analyses and their conclusions will be displayed, as well as limitations and future prospects.

2 Background

In the research on youth suicide, communities and specific institutions (eg, schools, universities, psychiatric units, and youth offender units) are two important geospatial elements that are frequently mentioned [4]. In previous studies, the geospatial attributes of the community were often mentioned and associated analysis with adolescent suicides [6]. The school, as a specific institution, is linked to related research, usually only as a scene of suicide intervention plans, aimed at investigating and verifying the effectiveness of school-based suicide prevention interventions [7, 8]. Its geospatial attributes are not stand out in the research.

However, these previous plans tended to focus on providing one or several general school intervention plans, through the introduction of professionals to provide relevant training for teachers and other school staff [8]. Although these plans have both universal applicability and ease of operation, in actual applications, they often encounter realistic bottlenecks. The bottleneck comes from the universality of the plan. This universality has caused the lack of utilization of the school's unique geospatial factors, and therefore it is difficult or even not to formulate and provide localized plans and response methods for specific schools or regions. On the other hand, the implementation of school suicide intervention plans largely depends on the identification of students' suicidal tendency. If the relevant suicidal tendency can be successfully identified as soon as possible before the suicide behavior is implemented or completed, it is likely that more targeted suicide intervention can be implemented.

Suicide ideation, the major risk factor for both suicidal attempts and completed suicide, may be an important link in identifying suicidal tendencies and interfering with suicidal behavior [9]. Suicide ideation is also a main risk indicator that is used in clinical practice by Child and Adolescent Mental Health Services (CAMHS) professionals. Questions about suicide iteation are often mentioned in routine clinical practice, especially when the patient has shown depression, anxiety, difficulty in mood regulation, trauma or distress following difficult life experiences etc. [9]

Suicide ideation information could be documented in both structured files or unstructured files, risk assessment forms [10] and free-text in Electronic Health Records (EHRs) respectively [9]. In clinical practice, EHRs is widespread adopted in hospital care systems. To capture and extract suicide-related information from text within EHRs [11], NLP approaches are essential. By combining the NLP method with the machine learning approaches, we could attempt to model and predict patient suicide ideation from the content of those EHRs, anaysed at different levels of granularity [12].

3 Related Work

3.1 Suicidal Adolescents Identification

Commonly used NLP methods to identify suicide awareness from EHRs text include: determining the appearance of related words or concepts (such as the word suicide and its various variants), or combining context to determine the meaning mentioned.

However, it is often difficult to directly extract patient-level suicide awareness or risk from EHRs text. Each patient's EHR may include multiple or even multiple files, and each file may contain important clinical information [9]. The existing suicidal tendencies approaches are usually based on the mention level or document level [3].

Down et al. has previously developed a step-wise rule based NLP method that can identify and filter out negative examples [13]. This NLP method determines the label of the patient level by judging the label of the document level, which is based on the principle of majority decision of mention level (that is, if the patient has one or more documents marked as suicidal positive, the patient is also marked positive for suicidal behavior at the patient level). After training and testing, this method has a precision score of greater than 0.85 on document level.

Song et al. recently innovatively proposed a method of recognizing suicide awareness based on sentence level, thus avoiding the loss of context information caused by using mention level [3]. At the same time, considering that the mention of suicidal behavior may occur in a cross-sentence context, he uses the Context-LSTM-CNN [12] algorithm to consider both the inter-sentential and intra-sentential contexts. The accuracy rates at the sentence level, document level and patient level have been significantly improved, respectively 98.71, 85.64, 85.31 . However, it only used the model in a corpus of Autism Spectrum Disorder (ASD) patients, and did not apply it in the generally and broader corpus.

3.2 Geospatial in Suicidal

Existing studies on the geographic distribution of suicide usually point out that suicide incidents tend to cluster. According to an 18-year spatiotemporal model analysis of New Zealand found that 1.3% of suicides may occur in clusters [14]. McKenzie et al. pointed out that suicides tend to occur in groups among young people, and the clustering behavior of suicide is more common among young people (<25 years old) than adults [4]. Especially in certain institutions (such as schools, universities, etc.), the number of suicide incidents is often more than expected, and involves geographically distributed link plots. Once a suicide cluster is formed, the impact on the community and specific places will be profound. Suicide cluster means a cluster of suicide events, usually in time, place, or both, more suicide events than expected, usually including three or more deaths [15].

In a study using two clustering methods, Space-time clustering and Space-time-method clustering, for 2741 suicide events, two significant clustering phenomena were found in patients with mental illness [16]. In another study of young Australian suicide clusters, a comparison research of three different cluster detection methods have been made(The scan statistic, Coronar inquests into suicide clusters, Descriptive network analysis). All found relatively high-risk spatiotemporal clusters in Australia [17].

3.3 Suicidal and School

In past studies linking adolescent suicide behaviors to schools, most of them have focused on the impact of school violence or bullying in schools on adolescent suicide behaviors [18], or research on school-based suicide prevention/ intervention‘[19, 20, 7]. This project attempts to study whether the geographic distribution of adolescent suicide rate is significant at the school level from a geospatial perspective, and the impact of student flow on the adolescent suicide rate. The use of predicted suicidal tendencies instead of the existing suicide records data is that the suicidal tendencies predicted by the deep learning model can significantly determine whether the patient is at risk of suicide, thus providing support for suicide prevention and intervention to a certain extent. On the other hand, by analyzing the data of the prediction of suicidal tendency at the geospatial level, we can better judge the multiple relationship between the suicidal tendency in geospatial and the school, so as to better provide support for suicide prevention at the school level.

4 Approach

4.1 Overview

In this paper, the deep neural network is used to identify suicide behaviors from EHRs to apply the suicide prevention annotations to either document or patient level, and then geospatial analysis is performed to correlate the annotation results with school data in London. In order to identify suicide ideation, a sentence level suicidal behavior classification approach based on a Context-LSTM-CNN [12] algorithm would be used.

This section will cover the C-LSTM-CNN algorithm followed in order to predict the suicidal adolescents from mental health records. It will then move to the variety of analysis methods used in the geospatial analysis, including global spatial autocorrelation, local spatial autocorrelation, and the cluster analysis. Finally, all data sets used in this research will be introduced.

4.2 C-LSTM-CNN

The Context-LSTM-CNN algorithm is shown in Figure 1. While doing the suicidal iteation identification, he architecture takes the left context, the focus sentence, the right context as inputs, to create both the intra- and inter- context for sentence. It is based on the following components: Word embedding, Bi-directional LSTM, Multiple window CNN, Context encoder, Softmax classifier.

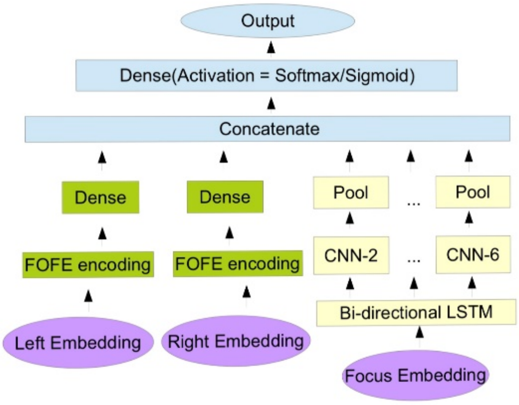


Figure 1: The C-LSTM-CNN algorithm

Every single words from the inputs will be transformed to vector space representation first, which using the Word2Vec(w2v) embedding model [21] , and then for the focus sentence, the Bi-directional LSTM [22] would be used to enrich the word vector representation with sentence level sequential information. The third part for input focus sentence is the CNN [23], extracting specific information from it.

As creating the Intra-context for the every focus sentence, the left and right context are going to be encoded by the FOFE encoder.Finally, a softmax layer takes both the inter- and intra- contexts ouput and combines them.

According Table 1, C-LSTM-CNN The achievement accuracy of the model at the three levels of sentence level, document level, and patient level are significantly higher than other models (SVM, MaxEnt, CNN only, LSTM only, LSTM-CNN), which achieves 85.31% accuracy on the patient level degree, having the best performance at all three different level [3].

Table 1: Four class mean accuracy

Model	Sentences	Document	Patient
SVM	85.89	61.53	63.70
MaxEnt	77.60	59.68	63.71
CNN only	97.98	80.81	80.81
LSTM only	98.12	81.92	80.17
LSTM-CNN	98.61	85.28	84.47
C-LSTM-CNN	98.71	85.64	85.31

4.3 Geospatial Autocorrelation

The geospatial analysis part will use the previously generated patient-level suicide training results combined with school data for analysis. Through the LSOA geographic location of each patient, combined with the data of student flow from LSOAs into each schools, the suicide tendency score of each primary school is simulated. This score represents the number of students who may have suicidal tendencies in the simulated school, and plays an important role in the geospatial and non-geospatial analysis of suicide and the school. In the specific analysis, some schools with a suicidal tendency score less than 1 are ignored, because this means that in the simulation results, there is less than one student who may have suicidal tendency in the school, and it can be determined that there is no suicide in the school. Tendency students.

According to the suicidal iteation score of each school, this article has done corresponding geospatial cluster analysis, global and local spatial autocorrelation, trying to explore the relationship between student suicidal iteation and other geospatial elements in the school.

4.3.1 Global Spatial Autocorrelation

Spatial autocorrelation is a significant research method used in spatial analysis and spatial regression analysis, which refers to the correlation between the attribute value of the research object and its spatial location. Spatial autocorrelation is often used to test whether there is a significant relationship between the attribute value of an element with it's neighbors‘. A positive correlation indicates that the attribute value change of a unit has the same changing trend as its neighboring spatial unit, and a negative correlation The opposite is true.

Global spatial autocorrelation is a description of attribute values for spatial characteristics in the whole region. It analyzes the spatial correlation and spatial variability of the entire region mainly by estimating Global spatial autocorrelation statistics (such as Global Moran's I, Global Geary's C, and Join Count). One of the most commonly used is Moran's I. The closer its value is to 1, the smaller the overall spatial difference will be. On the contrary, the closer the value is to -1, the greater the overall spatial difference will be.

The global Moran's I value is an aggregate statistic that describes only the average degree of spatial difference between all regions and surrounding regions. In the case of reducing the overall regional spatial differences, the local spatial differences may be enlarged. For the Global Moran's I statistics, the null hypothesis indicates that the analyzed attributes are randomly distributed among the elements in the study area. In other words, the spatial processes used to promote patterns of observations are random.

4.3.2 Local Spatial Autocorrelation

Based on the Local Indicators of Spatial Association (LISA) proposed by Anselin in 1995 [24], this article can reveal the spatial autocorrelation properties of the local area up to each spatial unit. LISA essentially decomposes Moran's I into various regional units, and for a certain spatial unit i can be expressed as Ii. If Ii is significantly greater than 0, it indicates that the spatial difference between area i and the surrounding area is significantly small; if Ii is significantly less than 0, it indicates that the spatial difference between area i and the surrounding area is significantly greater. The local spatial correlation can also be described by Moran scatter plot. Moran scatter chart is divided into 4 quadrants: A. The upper right quadrant HH, the spatial difference is small, the area itself and the surrounding level are higher; B. The lower left quadrant LL, the spatial difference is small, but the area itself and the surrounding level are lower ; C. The upper left quadrant HL, the spatial difference is large, the region's own level is higher, but the surrounding is lower; D. The lower right quadrant LH, the spatial difference is larger, the region's own level is lower, but the periphery is higher.

4.4 Spatial Clustering Analysis

Spatial clustering analysis is one of the important methods of spatial pattern recognition and spatial data mining. Specifically, it refers to dividing the objects in the spatial data set into classes composed of similar objects. Objects in the same category have a high degree of similarity, while objects in different categories have greater differences. As an unsupervised learning method, spatial clustering does not require any prior knowledge, such as pre-defined classes or labels with classes. However, the current spatial clustering analysis method has two biases. One is clustering based on the geographic coordinates of the spatial object, that is, only considering the spatial proximity of the object, without considering the similarity of the object attribute characteristics; the other is based on the traditional The cluster analysis method analyzes according to the attribute feature set, while ignoring the spatial proximity of the object. In this study, corresponding spatial clustering analysis was done for the two methods of geographic coordinates and geographic coordinates combining attribute characteristics, and the combination of the two was used to describe the spatial characteristics and spatial differences as completely as possible.

In this study, three clustering methods, K-Means, DBSCAN, and OPTICS were selected successively. After comparison and screening, the K-Means method was finally selected to perform spatial clustering analysis on the data.

4.4.1 K-Means

Due to the simplicity and efficiency of K-means clustering, it is the most famous clustering algorithm and also the most widely used clustering algorithm. The first step of the whole architecture is to divide the full data into K groups, randomly select K objects as the initial clustering centres.Following calculate the distance between each object and each initial clustering centre, and assign each object to the clustering centre closest to it. For each cluster centres, every single objected that assigned to it together represent a cluster. Each time a sample is allocated, the cluster centre would be recalculated based on the existing objects in the cluster until certain termination conditions are met. The termination condition can be that no (or minimum number) objects are reassigned to different clusters, no (or minimum number) cluster centres change again, and the sum of the squared errors is locally minimum.

The k value of K-Means needs to be specified by the user, but this data set is difficult to distinguish the number of clusters from the naked eye, and there are more noise points (outliers) in the spatial distribution.

4.4.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a representative density-based clustering algorithm. Different from partition and hierarchical clustering, it defines clustering as the maximum set of points connected by density, can divide regions with sufficient density into clustering, and can find any shape clustering in noisy spatial database.

DBSCAN has some huge advantages over K-means clustering algorithm. First, it doesn't require a preset group set at all. It also treats outliers as noise, which is different from the K-Means algorithm, put data points into a cluster even if they are very different. In addition, the algorithm can also find any size and shape of the cluster.

The main disadvantage of DBSCAN is that it will not perform as well as other algorithms when the cluster density changes. This is because when the density changes, the distance thresholds and minPoints used to determine adjacent points will also change. Because the distance threshold becomes difficult to estimate, this disadvantage can also occur in very high performance data.

4.4.3 OPTICS

The OPTICS (Ordering Points to Identify the Clustering Structure) clustering algorithm is a density-based clustering algorithm. The goal is to cluster data in space according to the density distribution. The idea is very similar to DBSCAN, but it is actually one of the DBSCAN algorithm. This effective extension mainly solves the problem of sensitivity to input parameters. However, unlike DBSCAN, OPTICS algorithm can obtain clusters of different densities. In other words, after OPTICS algorithm processing, it is theoretically possible to obtain clusters of arbitrary density. Because the output of the OPTICS algorithm is an ordered queue of samples, clusters of arbitrary density can be obtained from this queue.

It was first proposed by Peter J. Rousseeuw in 1986 [25], combines the two factors of cohesion and resolution. It can be used to evaluate the influence of different algorithms or different operation patterns of algorithms on clustering results based on the same raw data. Although OPTICS clustering removes discrete points well, its Silhouette Coefficient results show less than 0. Silhouette Coefficient, which is an evaluation method of clustering effect.

4.5 Statistical Data

4.5.1 ASD_AMIA2017

The 'ASD_AMIA2017' dataset is planned to be used to train the Context-LSTM-CNN model. It is actually a revised version of the corpus described in (Downs et al., 2017), which contains free-text documents from the EHRs of adolescent patients who have an ASD and have been referred to the South London and Maudsley NHS Foundation Trust (SLaM). These records were extracted from the Clinical Record Interactive Search (CRIS) system (Perera et al., 2016), a resource with de-identified EHRs which allows data retrieval for secondary data analysis, approved by the Oxfordshire Research Ethics Committee C (reference 08/H0606/71+5).

The corpus consists of documents from 499 ASD patients containing at least one predefined term related to suicidal behavior. Suicidal behavior mentions in each document were independently annotated by one of two domain experts. Mentions were loosely defined: the annotators were asked to label any explicit mention of suicidality in the text, marking each mention as suicidality risk positive (SR-Pos, i.e. a patient with suicidality risk), suicidality risk negative (SR-Neg, i.e. not a patient with suicidality risk) or uncertain. In total, 4,918 documents were annotated, containing 6697 SR-Pos, 3,701 SR-Neg and 1,097 uncertain mentions.

4.5.2 MEDINFO

In the test and running part, the MEDINFO data used is also from the CRIS database. The entire cohort consisted of all the patients aged 11-17 who had contact with CAMHS in SLaM during April 1, 2009 to March 31, 2016. All documents for these patients during this period were extracted (including incident records, letters, and free text) from different types of semi-structured assessments, resulting in 1,601,422 documents (totally 23,455 patients). Structured data was not included, and the project's ethical approval did not cover access to patient outcomes.

4.5.3 London Primary Schools

After identifying document level suicidal ideation, make relationship and analysis with the school data from the Great London Authority, which contains primary and secondary provision, including academies and free schools in London. This data set includes all the primary and secondary provision, including academies and free schools recorded in 2016 in the Greater London area. All schools are provided with complete school information, including school name, school type and phase, postcode, coordinates, etc. In addition to this, this dataset also includes a part of very special data that describes the student flow LSOAs into each schools for every single LSOA level region in 2016.

5 Results

5.1 Suicidal adolescents prediction result

After using the model trained according to the C-LSTM-CNN algorithm, it is calculated that on the patient level, a total of 5709 cases show a pos tendency on suicidal iteation. As shown in Figure 2, it shows the geographical distribution of all patients with pos tendency in London, and most of the patients are located in South London.

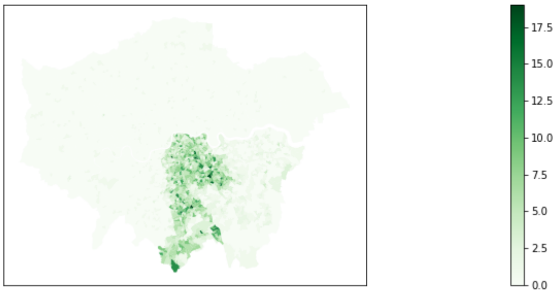


Figure 2: The distribution of suicidal ideation in LSOA level

5.2 Student flow from LSOA to Primary Schools

When the case is projected to the school with the student flow from LSOA to the primary school, the suicidal iteation score is assigned to the school according to the proportion of the number of people from LSOA to the school. Its projection on the map is shown in Figure 3, which is mapped to a total of 1,624 schools.

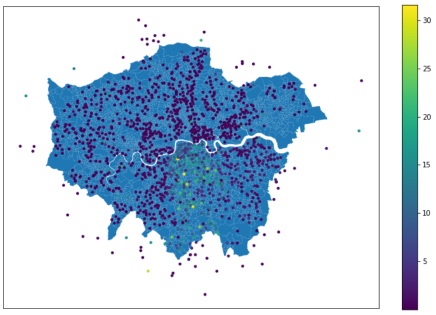
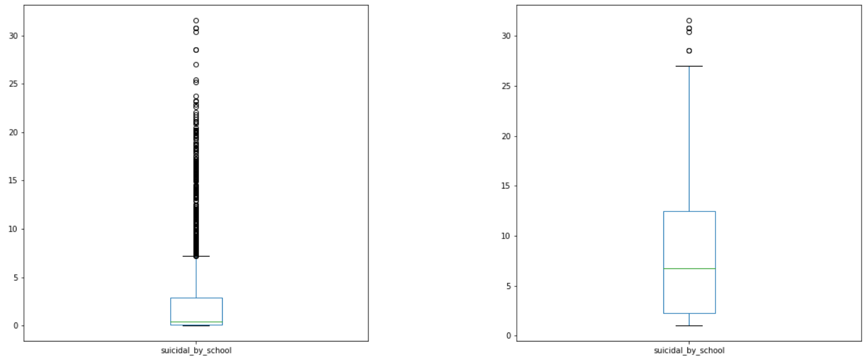


Figure 3: The projection of schools' suicidal score

The school with the highest attribute value is Rosendale Primary School in London, with a score of 31.572386. As shown in the Figure 4(a), the main distribution of the school's suicidal iteation score attribute value is between 0 and 5, and the median value is 0.41139299092246484. The data distribution is relatively skew, and there are a lot of outliers outside Q3+1.5IQR . Considering that for schools whose suicidal iteation score attribute value is less than 1, based on the existing data set, it is inferred that the number of internal students who have suicidal tendencies should be less than 1, which can be considered as non-suicidal students. After removing all schools with a suicidal iteation score less than 1, the box chart is shown in Figure 4(b). After filtering, 579 school data are retained, and the suicidal by school attribute value data distribution is more reasonable.



((a)) Boxplot before fliter ((b)) Boxplot after fliter

Figure 4: Boxplots before and after data flitering

5.3 Global Spatial Autocorrelation Results

The results of the spatial autocorrelation of the school's suicidal iteation score are shown in Table 2. Moran's I is 0.5151, z-value is 27.4616, p-value is 0.0000<0.0001, and the null hypothesis can be rejected at 99.99% level. , Indicating that the school's suicidal iteation score attribute value is similar in space with neighboring schools, and there is a more obvious trend of aggregation.

Table 2: Global Spatial Autocorrelation Results

Moran's I	0.4063
Expected Value	-0.0017
Z-value	12.9080
P-value	0.0000

Moran's I can also be represented by a Moran scatter plot, as shown in the Figure

5. The abscissa represents the attribute value of the school's suicidal by school, and the ordinate represents the spatial lag of the geospatial elements around each school point.

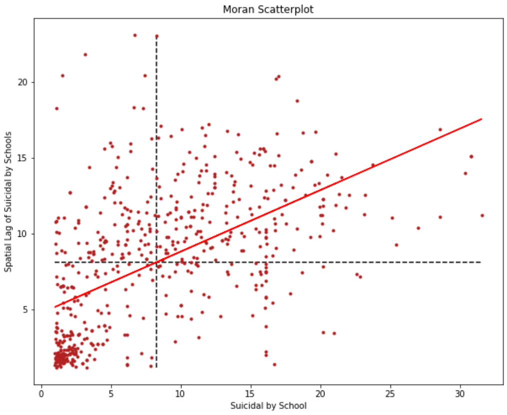


Figure 5: Moran's I scatterplot

5.4 Local spatial autocorrelation results

On the basis of global autocorrelation, further local autocorrelation analysis is performed on the suicidal iteation score attribute, and the result is shown in the Figure 6. A total of 33 school points are located in the HH quadrant, that is, their suicidal iteation scores are high, and the suicidal iteation scores of their surrounding schools are also high. Most of the 33 school points are located in the city of londo and Croydon, with 22 and 7 respectively. The other four areas have one each in Banstead, Bromley, Coulsdon and Tadworth.

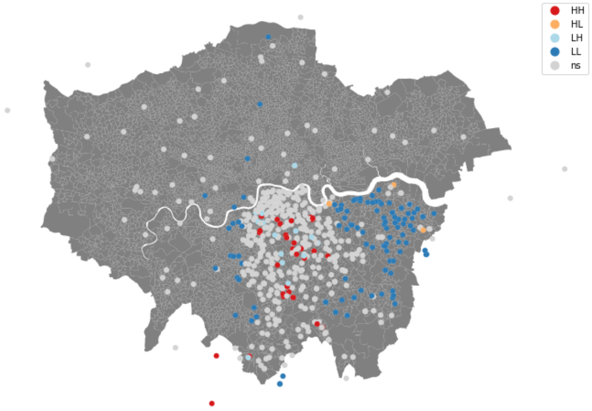


Figure 6: Local spatial autocorrelation results

5.5 Clustering Analysis Result

5.5.1 Geospatial Only

Given the value range of k from 3-9, a K-Means cluster analysis is performed for each value of k, and its cluster map projection and corresponding silhouette coefficient are output. As seen in the Figure 7, it is the cluster analysis result when k=6, and its average silhouette coefficient is 0.3837590070315731

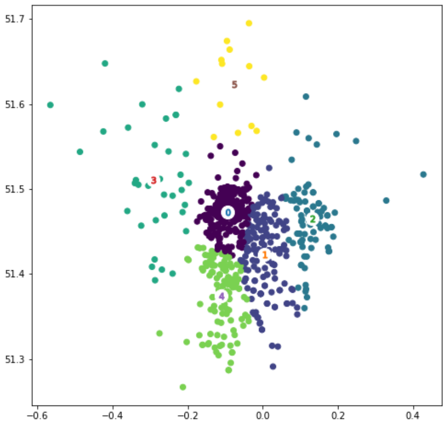


Figure 7: Geospatial only clustering analysis result

5.5.2 Geospatial with Attribute

After overlaying school suicide score attribute values with geographic coordinates, the school clustering situation has changed. As shown in the Table 3, when k is in the range of 3-9, the output average silhouette coefficient all reach about 0.6, indicating that the clustering results are better. The samples in the same cluster are dense enough, but different clusters The samples in between are relatively distant.

Table 3: Average Silhouette Coefficient for k in 3-9

K-value	Average Silhouette Coefficient
3	0.6240196858288998
4	0.6143371167636532
5	0.6056726110503141
6	0.6174713953648128
7	0.6066736855451018
8	0.5934344559292183
9	0.5919646067718155

Similarly, the case of k=6 is selected as an example, and the clustering situation with superimposed space and attributes is shown in the Figure 8. The projection of the clustering situation on the geographic space is more complicated, and the centers of each cluster are relatively concentrated.

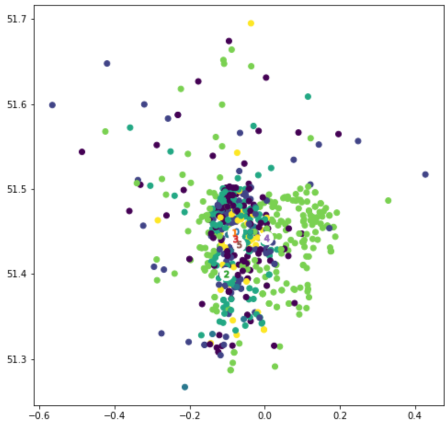


Figure 8: Geospatial and attitude clustering analysis result

6 Conclusion

6.1 Conclusion

The focus of this study is base on the previous studies among young people suicide clustering, and try to study whether the suicide tendency or suicide iteations of adolescents tends to gather at the level of special institutions (schools). The suicide tendency data of adolescents comes from the extraction of information in mental health records based on deep learning natural language processing methods. The results show that at the school level, the suicidal tendencies of adolescents have a high degree of spatial autocorrelation, and exact clustering phenomena are observed. The occurrence of clustering indicates that the school's suicide tendency index is not random, but highly correlated. Like suicide clusters, suicide clusters can also have a profound impact on communities or specific institutions. The suicidal tendency of young people is likely to influence each other and spread through geographic links.

At the same time, based on the results of local spatial autocorrelation, several schools with high suicidal tendency were identified, and both the school-level suicide awareness scores and the spatial lag in the surrounding geographic environment had high levels (Table 4).

6.2 Limitations

In this study, the EHRs text used to judge suicidal consciousness among teenagers comes from SLaM, so in the output, you can see that most of the patients whose suicidal tendency is marked as positive, as well as schools, are located near South London, while others in London The regional data is missing.Which has brought certain difficulties and obstacles to the cluster analysis of London as a whole, because it is impossible to determine what is the real level of the suicidal tendency of schools judged to be low suicidal in this study (especially schools located outside South London).

In addition, the spatial regression analysis did not perform as expected due to the limitations of the school data set. Because the characteristic attributes of schools, such as school type, school area, gender type, etc., are categorical attributes rather than statistical data, which limits the process of spatial regression analysis.

6.3 Future Work

Once the clustering phenomenon of suicide tends to occur, it shows that it does not occur randomly by chance, and there must be related reasons behind it. In addition to spatial autocorrelation, there may be other social phenomenon indicators related to it, such as social deprivation, regional economic development level, etc. Of the 33 high-risk schools in the HH quadrant found in this study, 22 of which are located in the city of london, does it mean that high levels of economic development may increase the suicidal tendency of teenagers? In the future, further analysis and research can be done by combining with other geospatial data sets. In addition, this study identified some clusters of suicide tendency through clustering and local spatial autocorrelation. In the future school-based suicide intervention and prevention, more targeted measures may be implemented. , Based on two different levels of individual and school. At the individual level, interventions and treatment measures for specific populations are provided for young people with suicidal tendencies. At the school level, schools with high suicidal tendencies focus on school-based suicide intervention programs.

Table 4: Schools in the HH quadrant

suicidal_by_school	SCHOOL_NAM	TYPE	TOWN
15.67973	St Anne's Catholic Primary School	Voluntary Aided School	Banstead
16.04619	Downderry Primary School	Community School	Bromley
18.76446	Smitham Primary School	Community School	Coulsdon
8.277388	Robert Fitzroy Academy	Academy Sponsor Led	Croydon
8.59995	Good Shepherd Catholic Primary School	Voluntary Aided School	Croydon
11.72924	Wolsey Infant School	Community School	Croydon
15.93509	Oasis Academy Shirley Park	Academy Sponsor Led	Croydon
15.93509	Oasis Academy Shirley Park	Academy Sponsor Led	Croydon
16.81012	Davidson Primary School	Community School	Croydon
23.748	Kingsley Primary School	Community School	Croydon
8.367386	Horniman Primary School	Community School	London
9.303283	Sudbourne Primary School	Community School	London
10.09206	Perrymount Primary School	Community School	London
10.62454	Our Lady and St Philip Neri Roman Catholic Primary School	Voluntary Aided School	London
11.11949	Riverside Primary School	Community School	London
11.52537	St Bartholomews's Church of England Primary School	Voluntary Aided School	London
11.74396	Ashmead Primary School	Community School	London
12.03055	Heber Primary School	Community School	London
12.42064	St Mary Magdalene Church of England Primary School	Voluntary Aided School	London
13.36051	St Anthony's Catholic Primary School	Voluntary Aided School	London
14.12624	Eliot Bank Primary School	Community School	London
15.27148	Lucas Vale Primary School	Community School	London
15.59769	St Stephen's Church of England Primary School	Voluntary Aided School	London
16.6921	Jubilee Primary School	Community School	London
16.91603	Dog Kennel Hill School	Community School	London
16.98541	South Norwood Primary School	Community School	London
16.99711	St Thomas Becket Catholic Primary School	Academy Converter	London
18.36211	Athelney Primary School	Community School	London
19.6846	Lyndhurst Primary School	Community School	London
21.08203	Kelvin Grove Primary School	Community School	London
30.7876	Durand Academy	Academy Converter	London
30.7876	Durand Academy	Academy Converter	London
28.55234	Kingswood Primary School	Community School	Tadworth

References

[1] WHO, "Suicide fact sheet."

[2] K. Hawton, H. Bergen, N. Kapur, J. Cooper, S. Steeg, J. Ness, and K. Waters, "Repetition of self-harm and suicide following self-harm in children and adolescents: Findings from the multicentre study of self-harm in england," *Journal of child psychology and psychiatry*, vol. 53, no. 12, pp. 1212–1219, 2012.

[3] X. Song, J. Downs, S. Velupillai, R. Holden, M. Kikoler, K. Bontcheva, R. Dutta, and A. Roberts, "Using deep neural networks with intra-and inter-sentence context to classify suicidal behaviour," in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1303–1310, 2020.

[4] K. Hawton, N. T. Hill, M. Gould, A. John, K. Lascelles, and J. Robinson, "Clustering of suicides in children and adolescents," *The Lancet Child & Adolescent Health*, vol. 4, no. 1, pp. 58–67, 2020.

[5] M. S. Gould, S. Wallenstein, M. H. Kleinman, P. O'Carroll, and J. Mercy, "Suicide clusters: an examination of age-specific effects.," *American Journal of Public Health*, vol. 80, no. 2, pp. 211–212, 1990.

[6] A. John, K. Hawton, D. Gunnell, K. Lloyd, J. Scourfield, P. A. Jones, A. Luce, A. Marchant, S. Platt, S. Price, *et al.*, "Newspaper reporting on a cluster of suicides in the uk," *Crisis*, 2016.

[7] J. Kalafat, "School approaches to youth suicide prevention," *American Behavioral Scientist*, vol. 46, no. 9, pp. 1211–1223, 2003.

[8] D. Wasserman, C. W. Hoven, C. Wasserman, M. Wall, R. Eisenberg, G. Hadlaczky, I. Kelleher, M. Sarchiapone, A. Apter, J. Balazs, *et al.*, "School-based suicide prevention programmes: the seyle cluster-randomised, controlled trial," *The Lancet*, vol. 385, no. 9977, pp. 1536–1544, 2015.

[9] L. Ohno-Machado and B. Séroussi, "Identifying suicidal adolescents from mental health records using natural language processing," in *MEDINFO 2019: Health and Wellbeing e-Networks for All: Proceedings of the 17th World Congress on Medical and Health Informatics*, vol. 264, p. 413, IOS Press, 2019.

[10] C. J. Hawley, B. Littlechild, T. Sivakumaran, H. Sender, T. M. Gale, and K. J. Wilson, "Structure and content of risk assessment proformas in mental healthcare," *Journal of mental health*, vol. 15, no. 4, pp. 437–448, 2006.

[11] K. Haerian, H. Salmasian, and C. Friedman, "Methods for identifying suicide or suicidal ideation in ehrs," in *AMIA annual symposium proceedings*, vol. 2012, p. 1244, American Medical Informatics Association, 2012.

[12] X. Song, J. Petrak, and A. Roberts, "A deep neural network sentence level classification method with context information," *arXiv preprint arXiv:1809.00934*, 2018.

[13] J. Downs, S. Velupillai, G. George, R. Holden, M. Kikoler, H. Dean, A. Fernandes, and R. Dutta, "Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records," in *AMIA annual symposium proceedings*, vol. 2017, p. 641, American Medical Informatics Association, 2017.

[14] G. Larkin, A. L. Beautrais, and H. Xu, "Geospatial mapping of suicide clusters in cantebury new zealand," *Injury prevention*, vol. 16, no. Suppl 1, pp. A258–A258, 2010.

[15] C. Niedzwiedz, C. Haw, K. Hawton, and S. Platt, "The definition and epidemiology of clusters of suicidal behavior: a systematic review," *Suicide and Life-Threatening Behavior*, vol. 44, no. 5, pp. 569–581, 2014.

[16] N. McKenzie, S. Landau, N. Kapur, J. Meehan, J. Robinson, H. Bickley, R. Parsons, and L. Appleby, "Clustering of suicides among people with mental illness," *The British Journal of Psychiatry*, vol. 187, no. 5, pp. 476–480, 2005.

[17] N. Hill, L. Too, M. Spittal, and J. Robinson, "Understanding the characteristics and mechanisms underlying suicide clusters in australian youth: a comparison of cluster detection methods," *Epidemiology and psychiatric sciences*, vol. 29, 2020.

[18] S. Bauman, R. B. Toomey, and J. L. Walker, "Associations among bullying, cyberbullying, and suicide in high school students," *Journal of adolescence*, vol. 36, no. 2, pp. 341–350, 2013.

[19] C. Katz, S.-L. Bolton, L. Y. Katz, C. Isaak, T. Tilston-Jones, J. Sareen, and S. C. S. P. Team, "A systematic review of school-based suicide prevention programs," *Depression and anxiety*, vol. 30, no. 10, pp. 1030–1045, 2013.

[20] J. Kalafat and M. Elias, "An evaluation of a school-based suicide awareness intervention," *Suicide and Life-Threatening Behavior*, vol. 24, no. 3, pp. 224–233, 1994.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[24] L. Anselin, "Local indicators of spatial association—lisa," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[25] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

A Spatial Lag for suicidal iteation score

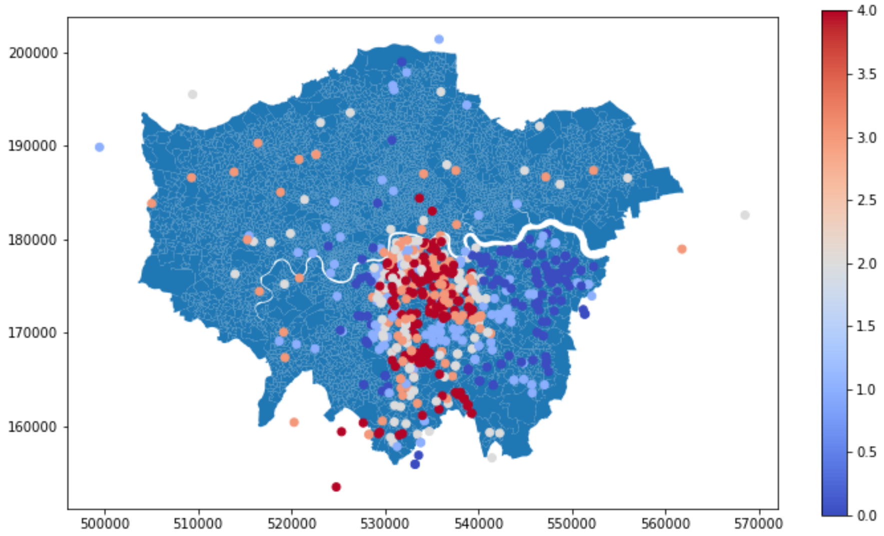


Figure 9: Spatial Lag

B K-Means cluster for geospatial only

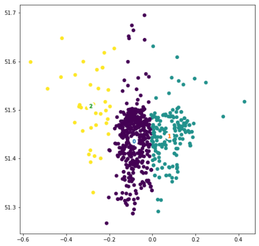


Figure 10: Geospatial K-Means for K=3

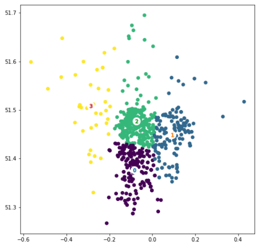


Figure 11: Geospatial K-Means for K=4

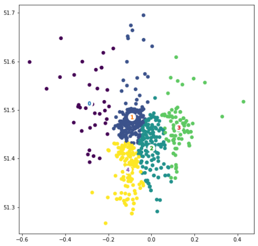


Figure 12: Geospatial K-Means for K=5

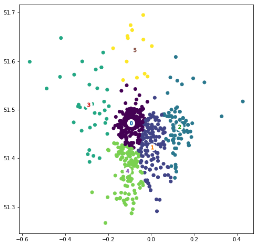


Figure 13: Geospatial K-Means for K=6

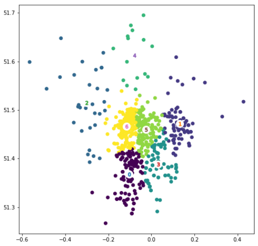


Figure 14: Geospatial Suicidal K-Means for K=7

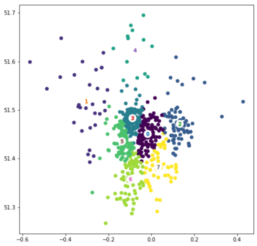


Figure 15: Geospatial K-Means for K=8

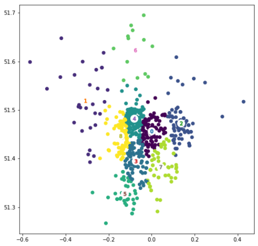


Figure 16: Geospatial K-Means for K=9

C K-Means cluster for geospatial with suicide iteation

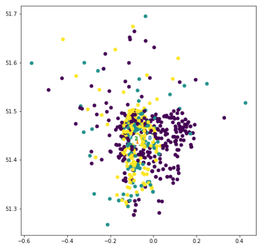


Figure 17: Geospatial Suicidal K-Means for K=3

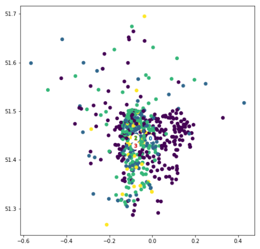


Figure 18: Geospatial Suicidal K-Means for K=4

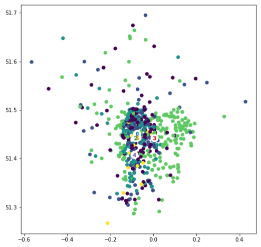


Figure 19: Geospatial Suicidal K-Means for K=5

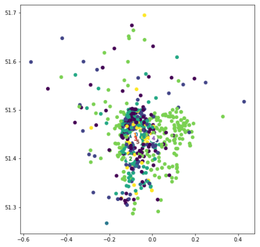


Figure 20: Geospatial Suicidal K-Means for K=6

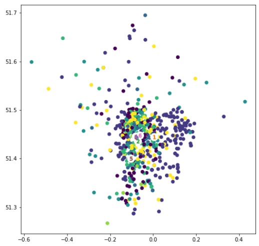


Figure 21: Geospatial Suicidal K-Means for K=7

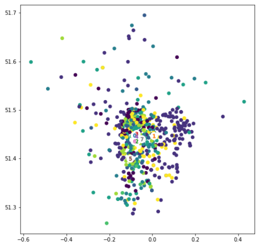


Figure 22: Geospatial Suicidal K-Means for K=8

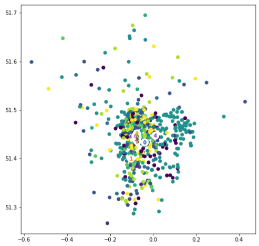


Figure 23: Geospatial Suicidal K-Means for K=9