Vol. 1, No. 1 | Jan – June 2017

Spatial Data Analysis: Recommendations for Educational Infrastructure in Sindh

Abdul Aziz Ansari, M. Abdul Rehman, Ahmad Waqas, Shafaq Siddiqui
Department of Computer Science, Sukkur IBA, Pakistan
aziz.mscs2013@iba-suk.edu.pk , rehman@iba-suk.edu.pk , ahmad.waqas@iba-suk.edu.pk, shafaq.siraj@iba-suk.edu.pk

Abstract

Analysing the Education infrastructure has become a crucial activity in imparting quality teaching and resources to students. Facilitations required in improving current education status and future schools is an important analytical component. This is best achieved through a Geographical Information System (GIS) analysis of the spatial distribution of schools. In this work, we will execute GIS Analytics on the rural and urban school distributions in Sindh, Pakistan. Using a reliable dataset collected from an international survey team, GIS analysis is done with respect to: 1) school locations, 2) school facilities (water, sanitation, class rooms etc.) and 3) student's results. We will carry out analysis at district level by presenting several spatial results. Correlational analysis of highly influential factors, which may impact the educational performance will generate recommendations for planning and development in weak areas which will provide useful insights regarding effective utilization of resources and new locations to build future schools. The time series analysis will predict the future results which may be witnessed through keen observations and data collections.

Keywords: Spatial analytics, Data Analytics, Education, GIS.

1. Introduction

Education is highly significant element for a developing country like Pakistan. Keeping this fact in perspective, Government of Pakistan has allocated sufficient amount of budget for improving education to Grass-root level. Standardized Achievement Test (SAT) is a reform initiative, a very timely and needful strategy to explore the dynamics of student learning in Sindh province. 'World Bank' also recognizes the effectiveness of SAT. Project SAT focuses on attitudinal changes in teachers and students effective learning influenced by the environment and infrastructure provided on regional basis. This study presents the analysis of data collected from SAT project. According to several reports including SAT-I, II and III, the quality of education throughout the Sindh province is alarming. According to SAT-II report, a test was conducted of the students of class V and VIII in three subjects, i.e. Science, Language and Math. SAT-II results show that the overall average score in all subjects is below 30% in all regions of Sindh province, which definitely a crucial situation. There may be several reasons of such failure, like teachers, physical infrastructure, language problems or the locations of schools.

ASER National Report 2015 indicates a critical education status in Sindh rural areas. The report indicates that less than 40% students capable of reading, writing stories in Sindhi, Urdu or English and doing basic mathematical operations [1]. ASER 2014 Sindh rural report shows alarming situations in various aspects of its education. Table 1.1 shows statistics gathered from ASER 2014 Sindh Rural Report [2]

Table 1: Learning Levels

LEARNING LEVELS (CLASS 5)			
English	24% can read sentences in English		
Urdu/Sindhi	41% can read story in Urdu/Sindh		
Arithmetic	31% can do 2-Digit division in arithmetic		
FACILITIES AVAILABLE FOR GOVERNMENT PRIMARY SCHOOLS			
Funds: 26%	Useable Water: 59%	Boundary Wall: 64%	Useable Toilets:48%

Reform Support Unit (RSU) also shows the statistics about the condition of education in sindh. Table 2 shows statistics of SAT 2014-15[3].

Table 2: Content Strand Based Scores ClassV

Subject	**Content Strand**	**Content Strand Average (%)**	**Subject & Overall Average (%)**	**Standard Deviation**
Language	***Reading***	54.16	32.81	18.6
	Writing	11.47		
Math	***Number & Operation***	18.70	18.22	12.78
	Measurement	37.74		
	Geometry	14.65		
	Information Handling	11.56		
Science	***Life Science***	14.76	15.26	11.04
	Physical Science	14.49		
	Earth & Space Science	28.46		
Overall Scores (%)			22.10	11.79

1.1. Geographical Information System

Geographical Information System (GIS) helps us visualize, analyse, interpret and understand data to reveal relationships and trends. According to Foorte, K.E and M.Lynch:"A geographic information system (or GIS) is a system designed to capture, store, manipulate, manage, and present spatial or geographical data" [4].In the beginning, use of GIS was aimed at the creation of maps only. The automation of paper based maps provided new idea of analysing data geographically using geometrical shapes and the database/linked data. This method was initiated by the Harvard Lab for Computer Graphics [5].

1.2. Quantum gis

QGIS (Quantum GIS) is stable open source geographic desktop application that provides efficient data viewing and analysis capabilities. Different countries and organization prefer GIS based analysis of available data that helps them in designing robust policies for the future of the country. Various independent international works have been carried out in order to infer the hidden factors that determine the progress of the education system in their particular country or region.

1.3. Geostatistics

Statistics is the science of producing facts and figures based on real/sample data by applying some analytical methods like finding averages, correlations, regression etc. This is an inferential approach to make decisions. The merger of GIS and Statistics came with new dimensions of analytics. In spatial/geo statistical analysis objects are represented by basic geographical symbols like lines, points and polygons. GIS presents spatial information to have independent analysis based on various features that highlight hidden patterns within data [5].

1.4. Time Series Analysis and forecasting:

Time series is a set of observed points x noted at an identified time t [6]. Plotted points express the growing or declining behaviour of data. The ordered series should be continuous in nature. Most of the time, A traditional time series is composed of two major components: Seasonal variation and Trends. Seasonal component includes analysis of growth or pattern in periods i.e. weekly, monthly, quarterly or yearly, while Trend component is based on linear increasing or decreasing trend [7]. Selection of method is based on the context and nature of data. Time Series Forecasting is a method used to predict the future data. Observed time series points x1,x2,…,xN can lead to the next possible trend xN+h where h (h for forecasting horizon) is the lead time. Most of the literature has divided forecasting in three general classes which may be used together in some situations.

- Judgmental forecasts based on subjective judgment or perception.
- Univariate methods based on heuristic data series having some linear trends.
- Multivariate methods based on some predictors [7]

According to the literature review [7] seasonality is considered as additive if it is not dependent of local mean and sum of the tables over years' values generally are stabilized to$\sum i_t = 0$. Seasonality is considered as multiplicative if size of the seasonal variation is related to the local mean and sum of the year's values can be normalized by modifying the averagei_t=1. Regression, Moving average and Exponential Smoothing are some of the popular forecasting methods.

1.5. Pearson Correlation Coefficient

According to National Council on Measurement in Education (NCME), correlation coefficient r is a numeric value that determines the statistical relationship or dependencies between two variable/attributes [8]. This can define the positive, negative or neutral effect of an attribute on other within the same cluster. A positive relation indicates that the increasing change in attribute A, affects the attribute B positively or increasingly. Negative relation indicates a negative or decreasing effect on attribute B when attribute A changes increasingly. No relation indicates that attribute A has no effect on attribute B. It measures dependencies of variables by value -1 <r< +1. Correlation coefficient r can be formulated as:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Where n is the number of data pairs and x and y are variables. Measures for correlation are as under:

- Correlation is said to be strong, if r≥±0.8
- Correlation is said to be weak, if r≤±0.4
- Correlation is said to be strong positive, if r is near to +1. If r=+1, then it is said to be a perfect positive correlation.
- Correlation is said to be strong negative, if r is near to -1. If r=-1, then it is said to be a perfect negative correlation.
- Correlation is said to be no-relation, if r is near to or is 0.

1.6. Univariate Forecasting Methods

As our data is not seasonal or multivariate, we must consider available methods to decide which would be appropriate to use. Available methods for univariate forecasting include Seasonal Moving Average (SMA), General/Simple Exponential Smoothing (SES) and Autoregressive Integrated Moving Average (ARIMA).

1.7. Seasonal Moving Average:

Moving average is a method to make time series data smooth by taking averages. Average is taken with or without weight values. This method can be applied to forecast data with seasonality. Forecast is predicted by observing average of last‘t’ terms/seasons. Average moves with respect to time. Moving averages are calculated as F=t1+t2+…+tn. Seasonality may occur in data for several reasons like climate conditions Moving average is a method to make time series data smooth by taking averages. Average is taken with or without weight values. This method can be applied to forecast data with seasonality. Forecast is predicted by observing average of last ‘t’ terms/seasons. Average moves with respect to time. Moving averages are calculated as F= (t1+t2+…+tn)/n to regulate the seasonality in time series. Seasonality defines episodic change in data which may occur for several reasons like climatic conditions, occasions, breaks etc. It spreads in different time periods like weekly, monthly, and quarterly or as factors discussed above. Seasonal variations can be expressed as a pattern or behaviour which leads towards future predictions.

1.8. General exponential smoothing

Exponential smoothing (also known as Simple or Single Exponential Smoothing) [9] is one of the empirical univariate forecasting methods [10]. Forecasting typically is based on smoothing averages of prior results. In a time series data, averages are supposed to be weighed by subsequent decaying patterns in averages. Unlike other univariate forecasting methods such as Moving Average (MA), Seasonal Moving Average (SMA) and Autoregressive Integrated Moving Average (ARIMA), General Exponential smoothing method is normally used to forecast such data where seasonality or trend is not present. Future value S_n+1is predicted by calculating weight average of the most recent results=aX_i+ (1-a) S_ (n-1). a is the constraint of level of the series which is considered as constant at local element of series and gradually changes over time [11]. The value of alpha can be selected by selecting the least Root Mean Squared Error (RMSE) where MSE = VARIANCE (errors) + (AVERAGE (errors)) ^2. Staring value (S0) is required in this method. Several methods have been proposed to calculate or to select the starting value. According to Gardner [11] although there is no significant empirical method of taking first value for forecast, but taking mean of data is the popular method while some consider first data segment in series as first value for forecasting. Gardner states exponential smoothing better as he found that accuracy of this method is amazingly accurate and easy to implement [11].

1.9. Autoregressive integrated moving average (ARIMA)

ARIMA is simplification of traditional Auto Regressive Moving Average (ARMA) model. It is used to predict future values from time series data. It is used mostly in the case where non-stationary process is observed which means there is a proof of change in mean and variance with respect to time. The integration of this model reduces the non-stationary process or non-seasonality. Box and Jenkins states three phases of ARIMA modelling: identification of time series properties of data, estimation of parameters of model and checking results of model with respect to hypothesis [10].

Rest of the paper is divided into 3 sections. Section 2 discusses the methodology. Section

3 demonstrates results and recommendations. Finally section 4 articulates conclusion and scope of future work.

2. Methodology

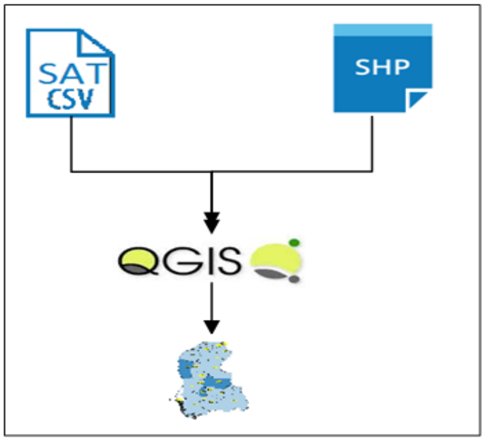


Figure 1: Architecture Diagram

Spatial Research framework shown in figure 1 shows the way this research work has been carried out. Following sections will briefly discuss each of the components in brief.

2.1. DATA PREPARATION:

Data preparation was a vital step of research. A chance of missing or unidentified record was present which was dealt. For preventing abnormal behaviours in analysis, data cleaning. Detected errors, altered outliers and filled missing data were performed. After removing the identified anomalies from the dataset, dataset was converted in a shape file. From shape file we extracted features which were appropriate to mapped against the spatial attributes of shape file.

2.2. Shape File:

Shape file is a simple, non-topological format for storing the geometric location and attribute information of geographic features [12]. We picked the map of Sindh in shape file with appropriate attributes to map data accordingly.

2.3. Data Mapping:

Collected data is in CSV (Comma Separated Value) format has been mapped with the features of shape file (.shp). The key spatial attributes for analysis were district attribute from the collected data set while it was the district spatial information attribute in the shape file. Figure 3.2 and 3.3 shows subject wise results of class 5 and 8 throughout Sindh province.

2.4. Feature Based Analysis:

Focus of this work is on finding influencing factors for education growth by analysing the results and facilities that are being provided in different schools across Sindh, Pakistan. So different attributes have been selected which are important parameters for highlighting the current status of facilities.

2.5. Data Visualization

For data visualization, we used SPSS and QGIS. Generated Scatter plots and histograms to visualize overall results as well as subject based results of class 5 and 8 students in Languages (Sindh/Urdu/English), Math and Science. Visual data stated the factual perspective of the problem.

2.6. Time Series Forecasting

Data compiled from last four year's (2013-2016) SAT results. Calculated overall averages for 23 districts of Sindh province, after compiling averages, identification of nature of time series data was the next step. Overall results were placed with years which concluded data as univariate by nature. Overall scores of 23 districts were mapped on four years, formulated as S_ij where S=Overall Scores, i=1, 2, 23 (districts) and j=1, 2, 3, 4(years). This formulation of data seems like panel data. As data is univariate and has no seasonality and no stable trend, we chose general exponential smoothing method for forecasting formulated asS_n=aX_i+ (1-a) S_(n-1). Alpha is the weight chosen by calculating the Root Mean Square Error

(RMSE) within values [13]. An optimal alpha value was selected from a number of calculation Alpha=Root (Mean Square Error), wherea≤0≤1. The table 3.3 shows RMSE for different values.

Table 1.3 Rmse at Different Alpha Value

Alpha	RMSE
0.1	2.61
0.2	2.69
0.3	2.75
0.4	2.80
0.5	2.85
0.6	2.91
0.7	2.99
0.8	3.09
0.9	3.21

RMSE 2.61 at Alpha 0.1 is the optimal weight in our case for smoothing forecasting [13]. After data preparation, the next step was to select first value for forecasting as it is the necessity of exponential smoothing method. Criteria for selecting first value were discussed earlier. We selected average of four years (2013, 2014, 2015, and 2016) data as first value for forecasting. Found predicted scores for 2017 to 2021 by using simple exponential forecasting method S_n=aX_i+(1-a) S_(n-1).

3. Results and Recommendations

Findings of analysis against geographical maps are discussed in this section. These findings lead us to propose some recommendations for authorities with some significant facts and figures. Most of the findings are based on factors derived from collected data.

3.1. Performance Influencing Factors:

This section includes discussion about some factors which may cause better performance and some assumptions which study has rejected. These factors were identified by significant statistical analysis on a number of attributes like scores, gender, medium, infrastructure, distance/location of schools, Teacher's information, student's family background, surveys, etc. All attributes were converted into numeric form for correlation analysis. Hypothetically by considering these factors, we can improve student's performance. Scores in three subjects Language, Math and Science generally influence overall performance and there is no need to find correlation. Table 4 shows correlation of different attributes with overall scores.

Table 4: Correlation of Attributes with Overall Average Scores

Factor	Correlation Coefficient
Working Male Teachers	-0.041**
Working Female Teachers	0.697**
Boundary Wall	0.590**
Toilet	0.102**
Drinking Water	0.508*
Electricity	0.51**
Computer Labs	0.571
Qualification Academic	0.401
Experience	0.518
Punctuality	0.555**
One task at a time	0.410*
Lesson plan	0.541*
Infrastructure	0.690
*. Correlation is significant at the 0.05 level (2-tailed).	
**. Correlation is significant at the 0.01 level (2-tailed)	

Results in table are clearly show positive, negative or no relation of attributes with overall performance. This rejects a common perception that spending more budgets only on building physical infrastructure leads towards better results. On behalf of our study we can say that including infrastructure, there are some important factors also, which may influence student performance. Cross-attribute analysis is another perspective of analysing correlation, through which we find positive correlation of Math subject with Science subject, which may conclude that by emphasizing student's cognitive level in math may increase results in science also. Table 5 and figure 4.1 states cross-subject correlations.

TABLE 5 CROSS-SUBJECT CORRELATIONS

Subject	**Correlation coefficient**		
	Language	***Math***	***Science***
Language	-	0.452	0.498
Math	0.452	-	0.586
Science	0.498	0.586	-

Study found that increasing the number of female teachers for lower classes, providing proper infrastructure with basic facilities (such as drinking water, electricity boundary wall and computer labs), appointing experienced teachers and implementation of lesson plans may cause increase in overall performance. After observing a number of school's data, it is found that most of the schools don't have sufficient facilities and required infrastructure. Infrastructure plays a vital role in school performance, as it is a significant influencing factor. These facilities include Electricity, Toilet, Drinking water and some other factors. Electricity is the key facility for students, administration and teaching staff in order to run all learning processes smoothly within a school. Figure 2 depicts the electricity facility available in all districts of Sindh, Pakistan. The District with green is the one in which majority of schools do not have electricity facility available. On other hand the districts with blue Color are those, schools which are availing electricity facility.

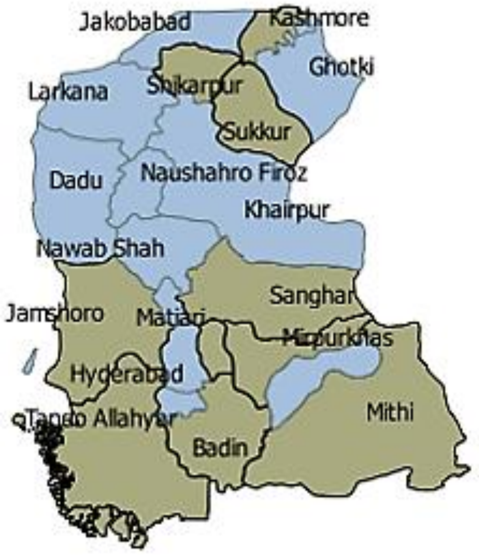


Figure. 2: Electricity Facility

Drinking Water is the most important facility for human being to be availed. The statistics in figure 3 shows that majority of schools in Ghotki, Dadu, Nawabshah even do not have facilities of drinking water. Absence of such basic need may result in the shortage of attendance which may lead to even catastrophic results. This facility must be provided in each school so that it may not harm the education system in these districts.

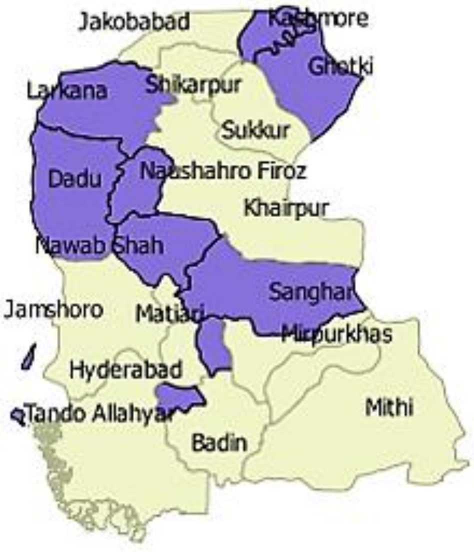


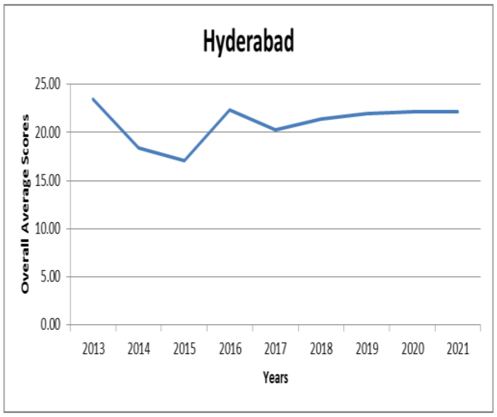
Figure. . 3 Drinking Water Facility

3.2. Time series forecasting

Table 5: Class 5 Overall Predicted Scores

District	2013	2014	2015	2016	2017	2018	2019	2020	2021
Badin	18.83	23.31	25.39	27.73	23.94	23.73	23.79	23.81	23.81
Dadu	17.88	20.88	23.86	26.08	22.30	22.09	22.15	22.17	22.17
Hyderabad	23.51	18.35	17.08	22.38	20.31	21.43	21.97	22.13	22.17
Jamshoro	22.48	21.30	23.37	27.94	23.86	22.77	22.34	22.21	22.18
Matiari	19.05	17.87	24.20	26.92	22.15	22.01	22.13	22.16	22.17
Shaheed Benazirabad	20.80	24.30	24.42	27.52	24.35	22.96	22.39	22.22	22.18
Tando Allah Yar	20.59	23.16	22.17	26.40	23.16	22.49	22.26	22.20	22.18
Tando Muhammad Khan	17.18	23.98	24.93	25.30	22.96	22.37	22.22	22.19	22.18
Thatta	15.67	19.14	22.06	25.66	20.78	21.44	21.96	22.13	22.17
Karachi City	32.00	27.00	23.74	30.50	30.70	31.71	31.95	31.99	32.00
Jacobabad	13.86	18.86	21.47	28.48	20.88	21.41	21.95	22.13	22.17
Qambar Shahdadkot	14.54	18.43	19.57	24.39	19.37	20.87	21.81	22.09	22.16
Kashmore	14.81	18.98	22.02	24.09	20.11	21.18	21.89	22.11	22.16
Larkana	15.41	17.11	17.92	22.35	18.30	20.48	21.70	22.07	22.15
Shikarpur	15.52	15.49	17.17	23.81	18.12	20.38	21.67	22.06	22.15
Mirpur Khas	19.81	25.68	26.32	25.68	24.45	23.02	22.41	22.23	22.19
Sanghar	18.03	24.00	19.88	26.26	22.13	22.05	22.14	22.17	22.17
Tharparkar	19.05	26.79	24.17	29.33	24.96	23.18	22.45	22.24	22.19
Umerkot	22.19	27.91	25.62	28.47	26.12	23.71	22.60	22.27	22.19
Ghotki	14.27	17.93	22.36	25.47	20.18	21.17	21.89	22.11	22.16
Khairpur	21.94	19.49	23.03	23.24	21.96	22.05	22.14	22.17	22.17
Naushahro Feroze	17.05	18.52	20.06	29.14	21.37	21.65	22.02	22.14	22.17
Sukkur	19.01	16.86	17.67	22.64	19.10	20.85	21.81	22.09	22.16

In this section, predicted scores and maps are discussed. Forecasting based on statistical analysis shows an approximate view of coming 6 years at district level which displays an insight into each district as well as whole province for the future. Results from 2013 to 2021(where scores from years 2013 to 2016 are real and from 2017 to 2021 are predicted) were compiled in SPSS and MS Excel. By using General Exponential Forecasting method we predicted scores for next five years. These predicted results were then mapped through QGIS. Table 6 shows predicted overall scores of Class 5.



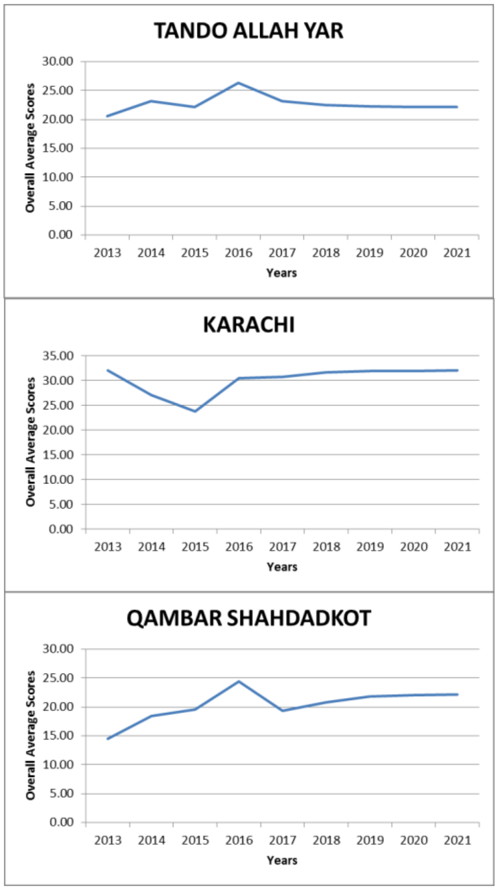
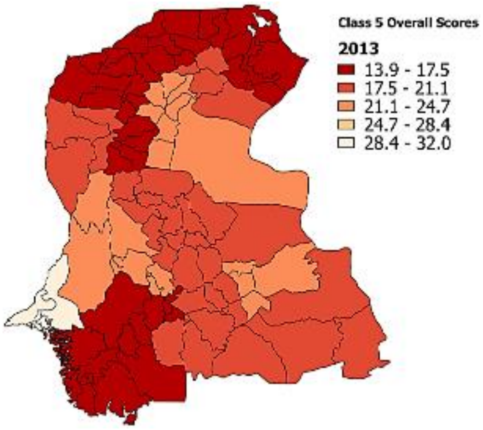
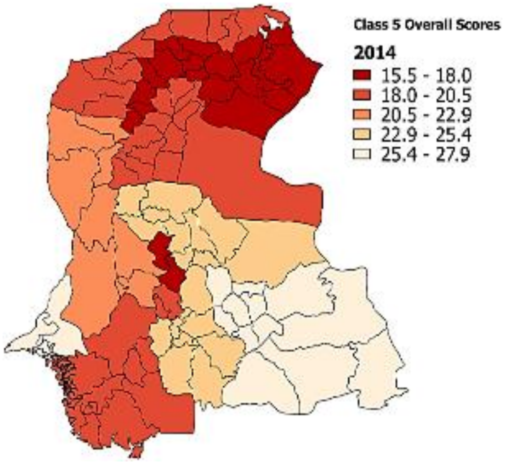


Figure. 4: Class 5 Overall Predicted Scores

Results shown above indicate crucial conditions for next 5 years, which may develop an assumption that the situation is getting worse. Class 5's results are showing a slowly increasing performance in some districts such as Karachi, Hyderabad, Qambar Shahdadkot, Larkana, Kashmore, Shikarpur, Ghotki and Sukkur. Above mentioned performance, As a matter of fact, it is still an alarming condition which must be considered as emergency, and needs to be tackled with immediate actions, whereas class 8's results indicate even worse condition. In class 8, most of the districts are showing a declining performance pattern, only a few districts like Karachi, Mirpur Khas and Sukkur display slow increasing performances, which still is a critical situation. Karachi is the only district which lies in the range of 25%-35% score, which if we consider as cross-match analysis among districts, is good, but still this is not overall a good result. Influencing factors should be considered for better performance. This study recommends that focus should also be on key cognitive development of subjects like Math and Science along with the provision of basic infrastructure throughout the province, which may produce better results. Figure 4.2.3and 4.2.4showing class 5 and 8 performance spatially with respect to time.





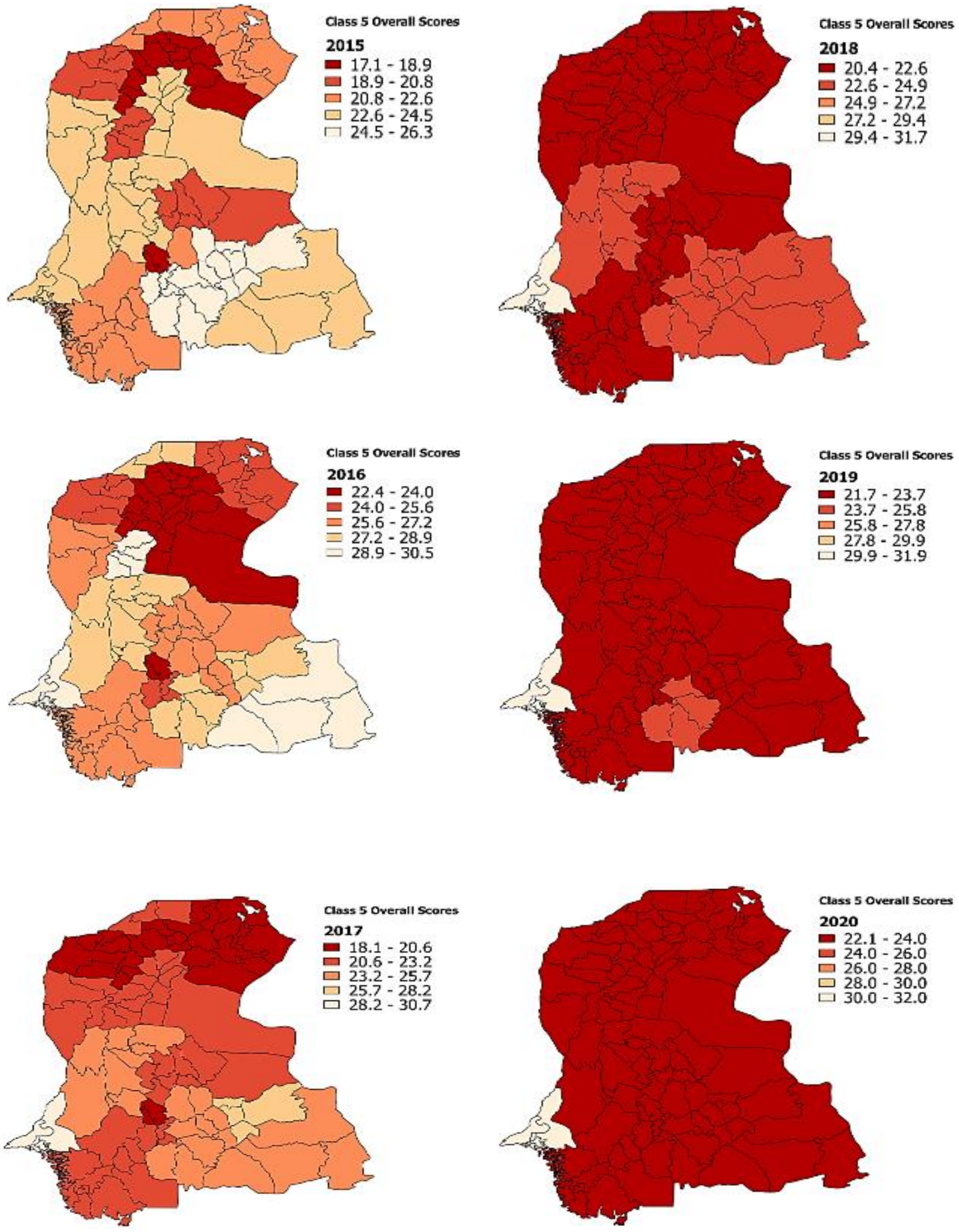


Figure.. 5: Class 5 Scores Map

4. Conclusion and Recommendations

Education is the most essential factor for development of a country. This factor is visualized graphically through geographical information system in all districts of Sindh, Pakistan. It has been observed in this research that even basic facilities like water, electricity and toilets are not available in various schools of Sindh, Pakistan. The task is to identify stimulating correlational factors that may cause raise in educational performance by analysing the datasets of schools, teachers and student's results, and predicting future image of situation in Sindh province. The analysis was carried out with respect to geographical locations at district level. There are some recommendations proposed by this study with the statistical support which should be considered to produce better results in future. Development and upgrading of infrastructure with basic facilities, appointment of well experienced faculty equipped with latest technology and techniques should be considered. Table 7 displays key influencing factors found in this study.

Table 7: Key Influencing Factors

Attribute	Correlation Coefficient
Working Female Teachers	0.697
Boundary Wall	0.590
Drinking Water	0.508
Electricity	0.510
Computer Labs	0.571
Experience	0.518
Lesson Plan	0.541
Infrastructure	0.690
Punctuality	0.555

Karachi, Hyderabad, Qambar Shahdadkot, Larkana, Kashmore, Shikarpur, Ghotki and Sukkur may produce better results in future. Future predictions are still alarming for critical situation of education performance at primary and secondary level.

5. Future work

Granularity is one of the key issues in any geo statistical analysis. Our study's granularity was at district level, which provides a brief overview of the case. Expanding the work to Taluka, UC and School level can shape a better insight. Socio-economic and political issues may influence the scenario that needs a comprehensive study. Collecting spatial information of all schools along with geographical coordinates may help develop a system in order to combine more aspects of education.

References

[1] ASER, ASER National Report 2015, 2015.

[2] ASER, Annual Status of Education Report (ASER) Scale & Scope, A.S. Scope, Editor 2014.

[3] RSU, Standardized Achievement Test (SAT)-III, E.a.L.D.G.o. Sindh, Editor 2014-15.

[4] Foote, K.E. and M. Lynch, Geographic information systems as an integrating technology: context, concepts, and definitions. Austin, University of Texas, 1996: p. 40-44.

[5] Burrough, P., Environmental and ecological statistics. Springer, 2001.

[6] Brockwell, P.J. and R.A. Davis, Time series: theory and methods2013: Springer Science & Business Media.

[7] Chatfield, C., Time-series forecasting2000: CRC Press.

[8] Education, N.C.o.M.i. Correlation Coeficient. Available from: http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorC.

[9] Wang, S., Exponential smoothing for forecasting and Bayesian validation of computer models, 2006, Georgia Institute of Technology.

[10] Liu, L.-M., et al., Forecasting and time series analysis using the SCA statistical system. Vol. 1. 1992: Scientific Computing Associates DeKalb, IL.

[11] Gardner, E.S., Exponential smoothing: The state of the art. Journal of forecasting, 1985. 4(1): p. 1-28.

[12] Gordon, I. and V. Monastiriotis, Education, location, education: a spatial analysis of English secondary school public examination results. Urban Studies, 2007. 44(7): p. 1203-1228.

[13] Ravinder, H.V., Determining The Optimal Values Of Exponential Smoothing Constants-Does Solver Really Work? American Journal of Business Education (Online), 2013. 6(3): p. 347.