

University of St Andrews
School of Computer Science



Big Data Analysis of School Data

Sabine Irene Weikert

110013687

Master of Science

Management and Information Technology

31st August 2012

Declaration

I hereby certify that this dissertation, which is approximately 13500 words in length, has been composed by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree. This project was conducted by me at The University of St Andrews from June 2012 to August 2012 towards fulfilment of the requirements of the University of St Andrews for the degree of MSc under the supervision of Dr. Adam Barker.

In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker, and to be made available on the World Wide Web. I retain the copyright in this work.

27.08.2012

Sabine Weikert

Acknowledgements

First I would like to thank my supervisor Dr Adam Barker for his assistance and guidance. His recommendations and suggestions have been very useful for the successful completion of this work.

A special thanks goes to Ralph Lucas, who inspired this project and enabled the data analysis by providing the necessary resources.

Last but not least I owe my deepest gratitude to my family for their unconditional support and encouragement throughout the year.

Table of Contents

Table of Contents.....	3
Table of Figures.....	4
Table of Tables.....	5
Abbreviation.....	6
Abstract.....	7
1. Introduction.....	8
2. Project Design.....	8
2.1. Initial Situation and Project Goal.....	8
2.2. Project Resources.....	9
2.2.1. Data.....	9
2.2.2. Tools.....	9
2.3. Project Scope.....	10
2.4. Project Methodology.....	10
3. Project Implementation.....	12
3.1. Data Understanding.....	12
3.1.1. Data Attributes.....	12
3.1.2. Data Visualization.....	16
3.1.3. Data Quality.....	29
3.1.4. Research Question Detailing.....	30
3.2. Data Selection.....	30
3.3. Data Analysis Model Development.....	31
3.3.1. Correlation between A-Level performance and school properties.....	31
3.3.2. Correlation between A-Level performance and UK Competitive Index.....	33
3.4. Data preparation.....	34
3.4.1. Data Cleaning.....	34
3.4.2. Data Integration.....	40
3.5. Data Analysis Model Deployment.....	44
3.5.1. Correlation between A-Level performance and school properties.....	44
3.5.2. Correlation between A-Level performance and UK Competitive Index.....	52
3.5.3. Summarization of results.....	61
4. Conclusion.....	61
Bibliography.....	63
Appendices.....	64

Table of Figures

Figure 1: Process Model	11
Figure 2: Data Model	13
Figure 3: Scatter plot A-Level points per pupil - Pupils in 6 th form 2003-2011.....	17
Figure 4: Scatter plot A-Level points per pupil - Pupil in 6 th form 2006-2011	18
Figure 5: Breakdown analysis - amount of data records and means 2006-2010	20
Figure 6: Boxplots A-Level points per pupil per gender 2006-2010.....	21
Figure 7: Boxplots A-Level points per pupil per school type 2006-2010	22
Figure 8: Boxplots A-Level points per pupil per school type and gender 2006-2010.....	22
Figure 9: Dispersion coefficients A-Level points per pupil per gender 2006-2010	23
Figure 10: Dispersion coefficients A-Level points per pupil per school type 2006-2010.....	24
Figure 11: Dispersion coefficients A-Level points per pupil per school type - gender.....	25
Figure 12: UKCI geographic map 2005, 2006 and 2008.....	26
Figure 13: UKCI geographic map 2009 and 2010	27
Figure 14: UKCI scatter plot 2005-2010	27
Figure 15: Geographical comparison A-Level points per pupil – UKCI 2010.....	28
Figure 16: Data records per cluster of the breakdown analysis	35
Figure 17: School data records per school type 2006-2010.....	36
Figure 18: School data records per gender 2006-2010	36
Figure 19: School data records per admission policy 2006-2010.....	37
Figure 20: School data records per year 2006-2010	37
Figure 21: School data records per school type and gender 2006-2010	38
Figure 22: School data records per school type and admission policy 2006-2010.....	38
Figure 23: School data records per gender and admission policy 2006-2010	39
Figure 24: School data records per school type, gender and admission policy 2006-2010.....	40
Figure 25: Number of schools included in integrated UKCI datasets	41
Figure 26: Number of local authority areas included in datasets.....	41
Figure 27: School types local authority areas CY local authority areas 2005-2010.....	42
Figure 28: IND local authority areas VA local authority areas 2005-2010	42
Figure 29: GFEC local authority areas FD local authority areas 2005-2010.....	43
Figure 30: Segmentation by geographical entities.....	44
Figure 31: Boxplots Pupils in 6 th form per school type 2006-2010.....	48
Figure 32: Boxplots Pupils in 6 th form per gender 2006-2010	49
Figure 33: Boxplots Pupils in 6 th form per gender and school type 2006-2010	50
Figure 34: School type clusters correlation A-Level points per pupil – UKCI.....	56
Figure 35: Geographic entity clusters Correlation A-Level points per pupil – UKCI.....	60
Figure 36: Summary of major findings.....	61
Figure 37: School data in regional context 2006 to 2009	65

Table of Tables

Table 1: Good Schools Guide data	9
Table 2: Attributes school league table – A-Levels	15
Table 3: Attributes school league table – A-Levels	16
Table 4: Quartiles UKCI 2005-2010.....	26
Table 5: Quartiles A-Level points per pupil 2010	28
Table 6: Quality assessment A-Level performance tables and UKCI dataset	30
Table 7: Data selection.....	31
Table 8: Correlation analysis methods.....	32
Table 9: Correlation A Level points per pupil – school properties.....	44
Table 10: Correlation Gender – A-Level points per pupil in different school types	45
Table 11: Correlation School type – A-Level points per pupil for different gender	45
Table 12: Correlation Admission policy – A-Level points per pupil per school types	46
Table 13: Correlation School type – A-Level points per pupil per admission policies.....	46
Table 14: Correlation Admission policy – A-Level points per pupil per gender	46
Table 15: Correlation Gender – A-Level points per pupil per admission policies	47
Table 16: Correlation Pupils in 6 th form – A-Level points per pupil per school type	47
Table 17: Correlation Pupils in 6 th form – A-Level points per pupil per gender.....	48
Table 18: Correlation Pupils in 6 th form – A-Level points per pupil school type - gender	49
Table 19: Correlation Pupils in 6 th form – A-Level points per pupil school type - gender	51
Table 20: Correlation A-Level points per pupil – school location 2005-2010	52
Table 21: Correlation A-Level points per pupil - UK Competitive Index.....	53
Table 22: Correlation CY A-Level points per pupil - UK Competitive Index	54
Table 23: Correlation IND A-Level points per pupil - UK Competitive Index.....	54
Table 24: Correlation FD A-Level points per pupil - UK Competitive Index.....	54
Table 25: Correlation VA A-Level points per pupil - UK Competitive Index	54
Table 26: Correlation GFEC A-Level points per pupil - UK Competitive Index	55
Table 27: Correlation CY A-Level points per pupil - UK Competitive Index	57
Table 28: Correlation IND A-Level points per pupil - UK Competitive Index.....	57
Table 29: Correlation London boroughs A-Level points per pupil - UKCI	58
Table 30: Correlation Metropolitan districts A-Level points per pupil - UKCI.....	58
Table 31: Correlation Unitary A-Level points per pupil - UKCI	58
Table 32: Correlation District A-Level points per pupil - UKCI.....	59
Table 33: Quartiles A-Level points per pupil 2006 to 2009	64

Abbreviation

A-Levels	Advanced Level
ADC	Art and Design College
AHC	Agricultural and Horticultural College
CTC	City Technology College
CY	Community School
GCSE	General Certificate of Secondary Education
GDP	Gross domestic product
GFEC	General Further Education College
ID	Identification
IND	Independent School
FD	Foundation School
KML	Keyhole Mark-up Language
KS 2	Key Stage 2
SFC	Sixth Form College
TC	Tertiary College
UKCI	United Kingdom Competitive Index
VA	Voluntary aided School
VC	Voluntary controlled School

Abstract

The research mission is a profound data analysis of large school and student datasets, with the objective of disclosing patterns and correlations among the data attributes. Using Tukey's (1977) exploratory approach to data analysis, an insight in the large datasets is gained, using different visualization techniques, as well as simple numerical statistics. Resulting, two research questions are stated, focusing on factors related to a school's A-Level performance: First, an analysis model, investigating on the correlation between the A-Level performance of schools and their admission policy, school type, gender and number of pupils in 6th form is developed and deployed. Subsequently, a correlation analysis between A-Level performance of schools and the UK Competitive Index, as indicator for the socio-economic situation of a local district is conducted. Correlations between a school's A-Level performance and its admission policy and number of pupils in 6th form are discovered. However, significant correlation between local A-Level results and the UK Competitive Index is not discovered, but a potential correlation within subgroups of the dataset, clustered by geographical entity. The results provide the grounds for further research on the correlation between educational as well as socio-economic factors and educational achievement.

1. Introduction

The research project “Big Data Analysis of School Data” is conducted in the context of a dissertation for the Master of Science Management and IT programme at the University of St. Andrews. The project mission is a profound analysis of large datasets of school and student data, with the objective of disclosing interesting patterns and correlations with respect to school performance. Therefore, different methods and techniques from the fields of exploratory, quantitative and visual data analysis as well as statistics are deployed. The research is assisted by the Good Schools Guide which provides access to the datasets and the business intelligence tool QlikView which is partly used to conduct the data analysis. In addition to the provided datasets, also datasets containing regional statistics are included, to extend the scope of research.

The first part of this report describes the framework of the research project in further detail, including an introduction of the project resources as well as the applied process model. The following core part of the dissertation concentrates on the actual data analysis. After a general understanding of the data is gained, a detailed research question is formulated. For the in depth investigation on the selected question, an appropriate data analysis model is deployed. The consequent results are presented and evaluated with regard to their significance and quality.

2. Project Design

2.1. Initial Situation and Project Goal

Good Schools Guide holds various data of English schools and students from different sources as for instance the National School League Tables and the National Pupil Database. The Good School Guide has already conducted a detailed analysis of the data, investigating on correlations and patterns that might be interesting for parents to help with their school choice (The Good Schools Guide, 2012). However the data analysis project presented in this dissertation is detached from previous research by Good Schools Data or other research institutions. The objective is rather to investigate the data with an unbiased analysis approach that is free of any presumptions, to potentially reveal correlations that have not yet been discovered. Furthermore, additional data records originated from reliable statistic offices may be integrated to draw a wider picture of national factors influencing the educational sector. In the end, the research results should be evaluated and visualized using innovative and

descriptive visualization techniques. The achievement of these objectives is enabled but also limited by the project resources which are discussed in the following chapter.

2.2. Project Resources

2.2.1. Data

The data provided by Good Schools Guide is specified in the table below, structured in the domains of school and pupil data.

Data	Description	Years	File format
School data	School Performance tables	A-Level, GCSE and KS2	2003 - 2011 .xls, .txt, .qvd ¹
	School Examination tables	A-Level, GCSE and KS2	2003 - 2011 .txt, .qvd
	School census		2007 - 2011 .txt, .qvd
	Ofsted reports		2005 - 2010 .xls, .txt, .qvd
	School Spending		2009 - 2010 .xls, .txt, .qvd
Pupil data	A-Level Pupil Examination tables	A-Level, GCSE and KS2	2003 - 2011 .txt, .qvd
	Pupil census		2007 - 2011 .txt, .qvd

Table 1: Good Schools Guide data

The detailed exploration of the data attributes and the evaluation concerning its quality and its usability and utility for the research project is described in Chapter 3.1, as first step of the actual data analysis.

2.2.2. Tools

QlikView

Good Schools Guide processes and analyses the data with the business intelligence tool QlikView. QlikView consolidates data from multiple files and supports different data analysis and visualization methods (QlikTech International AB, 2011). In this project, QlikView is used to integrate, search and discover the different datasets as it provides an intuitive interface for the analysis of big data. Furthermore, simple statistical calculations and the first general visualization of the data are conducted. However, to apply QlikView in order to do complex statistical calculations, elaborate skills and thorough knowledge is required. Due to the limited time scope of this research project, these skills could not be acquired.

¹ File format used by the business intelligence tool QlikView

Microsoft Excel 2010

Therefore, the detailed data visualization as well as the correlation coefficient calculations, shown in this dissertation, are conducted in Microsoft Excel 2010. This preference is due to the prior knowledge and experience of the researcher.

Google Fusion Tables

To visualize data analysis results, which include geographical components, Google Fusion Tables is used. This web application provides different data visualization techniques for tables, imported in .csv file format, for instance, the visualization of data on a map. In addition, imported tables can be filtered and merged, even with public tables, shared by other users (Google, 2012).

2.3. Project Scope

The dissertation project was scheduled over a period of three months. Regarding the tremendous amount of data to analyse, the time factor is the main limiting factor in this project, as it does not allow an in depth analysis of all provided data. Therefore the investigation is concentrated on a distinct research question – stated in Chapter 3.1.4 - that is selected after a general understanding of the data is gained. In addition to the time constraint, the access to the Good School Guide server and hence the data and QlikView was not given at any time.

The data analysis in this dissertation is not conducted by a sociologist or expert in the domain of the British school system but by a computer scientist. Therefore the investigation is exclusively based on statistical calculations and aims to be without bias concerning socio-economic assumptions and theories.

2.4. Project Methodology

The impartiality of the data analysis is promoted adopting the methodology of Tukey's (1977) exploratory data analysis. This approach is solely focused on the visual examination of data to find interesting research questions. Only for further investigation on the formulated research questions, statistical models are deployed. As advised by Berthold, Borgelt, Hoepfner and Klawoon (2010), the project is implemented based on the Cross Industry Standard Process for Data Mining.

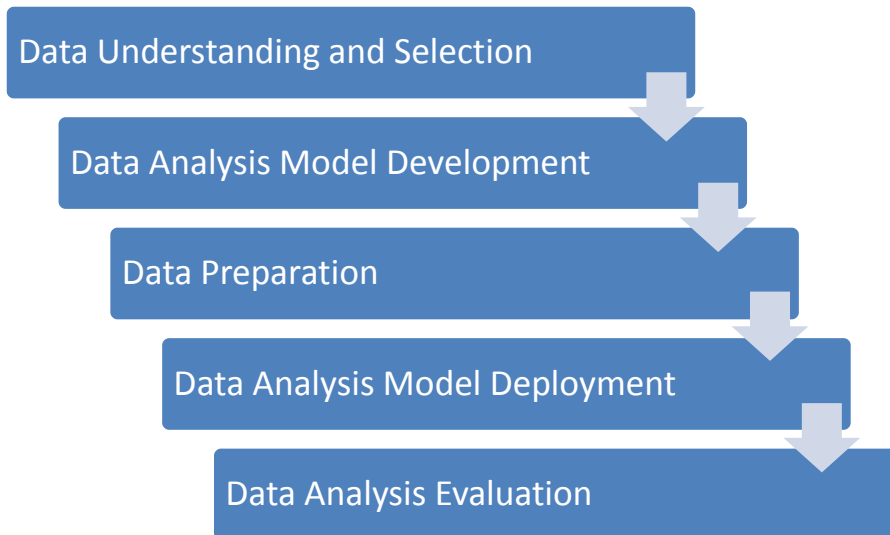


Figure 1: Process Model²

During the initial stage of the data analysis process, a general overview and understanding of the data is attained. Therefore, different visualization techniques and statistical measures are applied, to investigate on single data attributes and discover correlations between different data attributes. Furthermore the quality of the datasets and its attributes is evaluated. After the revision and detailing of the research question, the project stage is concluded with the selection of data for the analysis.

The second step of the process concentrates on the development of an appropriate data analysis model, to approach the revised research question. Under consideration of the desired structure of the analysis result, diverse statistical methods are selected and combined.

After the data analysis model is generated, the data preparation stage is set off. In respect of the data analysis methods chosen in the previous stage, the selected data is optimized in a cleaning and integration process.

Subsequently, the model is deployed and the results of the data analysis are summarized and visualized.

In the final process stage, the data analysis model and the consequent results are evaluated. The validity and coherence in regard to the research questions is assessed, reviewing each stage of the data analysis process. Subsequently, the informative value of the emergent results is rated.

²Based on Berthold, Borgelt, Hoepfner, & Klawoon (2010)

3. Project Implementation

This chapter documents the data analysis conducted in this dissertation project and is structured by the distinct process stages, explained in the previous section.

3.1. Data Understanding

3.1.1. Data Attributes

First, to gain a general overview of the data, the attributes of the school and pupil data are roughly explored. As summarization of this investigation, the data model shown in Figure 2 is created. This data model is the basis for the specification of the research area, as it reveals the relations between the data sets and the feasibility of their integration.

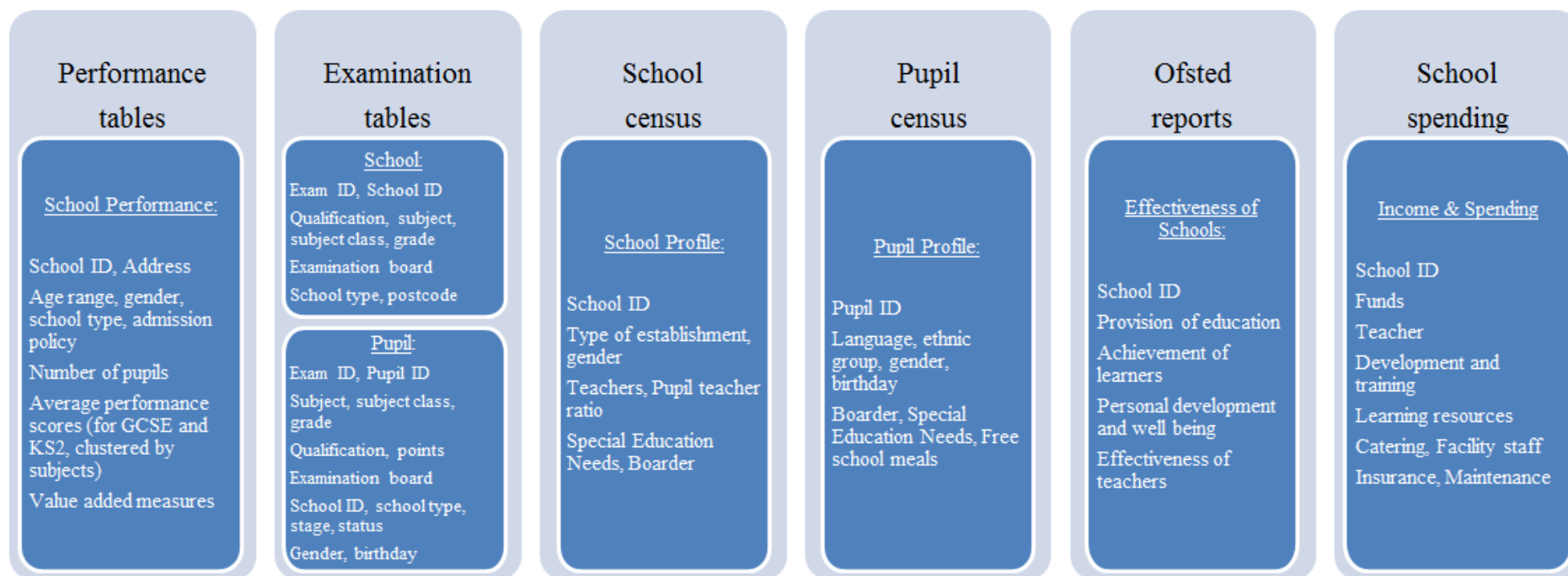


Figure 2: Data Model

The school and pupil datasets of the different years can separately be integrated by the primary key³ - School ID and Pupil ID, respectively. Furthermore the school data records can be set in relation to the pupil data by the School ID in the Pupil Examination tables.

³ Unique identifier for a data record

Due to the time constraint, an in depth analysis of all provided data is not feasible. In order to exclude data from further analysis, a rough specification of the research area is required at this project stage. Good Schools Guide has already analysed the data sets, exploring patterns related to the individual performance of pupils and schools in different subjects across the years. However, an investigation on national socio-economic factors influencing school performance has not been conducted yet. For further investigation on this research topic, regional datasets comprising appropriate information have to be included and set in relation to the existent datasets. For this purpose, the address attributes of the school league tables are capable of serving as foreign keys⁴. As the data attributes comprised in the school league tables also give information about the school's performance, they provide adequate information for further research. Hence, the pupil data, examination tables and Ofsted reports are excluded from further analysis.

To investigate on the relation between socio economic factors and school performance on national level, regional datasets comprising appropriate information have to be added. Therefore, the UK Competitive Index established by Robert Huggins (2003) is considered to act as an indicator for a region's socio-economic position on national level. This index is constructed based on a three factor model for measuring competitiveness. Input factors, identified as knowledge based business rate, economic activity rate and business density, contribute to the output of a region, specified as its productivity. Finally, this output, measured by GDP per capita, affects the outcomes of a region, specified as unemployment rate and earnings. Those factors are measured, using data from the Office of National Statistics, and weighted according to their correlation between one another. The UK Competitive Index is available for the years 2005 to 2010, excluding 2007.

Augmenting the data volume for the analysis with these regional datasets, the GCSE and KS 2 school league tables are excluded from further analysis as the investigation would exceed the time frame, set out for this project.

In the following tables the attributes of the A-Level performance tables and the UKCI are examined, implying only attributes available in the datasets of each year. Beside the description of the attributes, the appropriate data type is categorized. The data types are only quoted for numeric, ordinal and nominal data, as those have to be distinguished in the further data analysis. Numeric data is measured on a numeric scale whereas nominal data - also

⁴ Attribute that is primary key of another table and therefore enables cross-reference

known as categorical data - is assigned to a category, indicated by a name coding system. Ordinal data is similar to nominal data with the distinction that it is measured on an ordinal scale, giving the categories a sense of magnitude (Seale, 2004). This categorization is essential for the selection of analysis methods in Chapter 3.2.

A-Level performance tables

Attribute	Description	Data type
Year	2003 - 2010	Ordinal
Stage	A-Level	Ordinal
Country	England	Nominal
Mainstream_or_Special	Mainstream; Special	Nominal
DCSF_Reference_Number	Reference Number for the Department for Children, Schools and Families	
School_Name	Name of the institution	
School_Town	Address	
School_Postcode	Address	Nominal
School_Telephone	Telephone number	
School_Type	Academy; Agriculture and Horticulture College; Art, Design and Performing Arts College; City Technology College; Community School; Foundation School; General Further Education College; Independent School; Sixth Form College; Special School; Tertiary College; Voluntary aided School; Voluntary controlled School	Nominal
School_Admission_Policy	Comprehensive; Selective; Modern; Non-Selective	Nominal
School_6thForm_Gender	Mixed; Girls; Boys	Nominal
Age_Range	No fixed ordinal scale	
Pupils_in_6 th _Form	Number of pupils in 6 th form	Numeric
Pupils_in_Alevel_Tables	Pupils including in A-Level table	Numeric
Alevel_Points_per_Pupil	School's performance indicator	Numeric
Alevel_Points_per_Pupil_percentage_of_National	National comparison	Numeric
Alevel_Points_per_ALevel_or_Equivalent	School's performance indicator	Numeric
Alevel_Points_per_ALevel_or_Equivalent_percentage_of_National	National comparison	Numeric

Table 2: Attributes school league table – A-Levels

UK Competitive Index

Attribute	Description	Data type
Local authority area	Based on the Office for National statistics	Nominal
Districts	Based on the Office for National statistics	Nominal
UK Competitive Index	Local competitive indicator	Numeric
Index of Outcomes	Comprises earnings and unemployment rates	Numeric
Index of Outputs	GDP per capita	Numeric
Index of Inputs	Comprises knowledge based business rate, economic activity rate and business density	Numeric
Knowledge-based businesses	Percentage of all businesses	Numeric
% working age with nvq4+	Skill level equivalent to a bachelor's degree	Numeric
Business registrations	Per 10,000 inhabitants	Numeric
Businesses	Per 1,000 inhabitants	Numeric
Economic activity rate	Working age	Numeric
Employment rate	Working age	Numeric
GDP per capita	GDP per capita	Numeric
Productivity	Productivity	Numeric
Weekly median pay	Weekly median pay	Numeric
Claimant rate	Claimant rate	Numeric

Table 3: Attributes school league table – A-Levels

Subsequently, the preliminary research topic is selected:

Factors, influencing a school's performance in A-Levels

However, before a detailed research question is chosen, an in depth understanding of the data is gained, applying different data visualization techniques.

3.1.2. Data Visualization

Data visualization is the most significant technique of Tukey's (1977) exploratory approach to data analysis. It is not only used to visualize results but is a data analysis tool by itself, as it provides a summarization of the data (Mirkin, 2011). At this stage of the project, the A-Level school datasets and the UK Competitive Index datasets are visualized to explore patterns and evolve interesting research questions. Furthermore, distinct visualization techniques are used, to discover zero values and outliers. Those can be signs of bad data quality or, even if they

are correct, might be regarded separately throughout the data analysis, to ensure its coherence (Berthold, Borgelt, Hoepfner, & Klawoon, 2010).

A-Level School data

To begin with, the school datasets of the different years are merged together and analysed disregarding the year of their origin to conform to the unbiased approach of exploratory data analysis (Tukey, 1977). The following diagrams summarize the data included in the A-Level school league tables, focussing on the attribute *A-Level points per pupil* as indicator of the school's performance. First, the scatter plot in Figure 3 is created, to illustrate the value distribution of the attribute and highlight missing values and outliers. As a second numeric dimension is necessary for this visualization technique, the *A-Level points per pupil* are opposed to the corresponding value of the attribute *Pupils in 6th form*. The value distribution of the A-Level points clearly exhibits two accumulations, one between 100 and 500 points per pupil and the other between 400 and 1000. Furthermore a not negligible amount of null values is revealed.

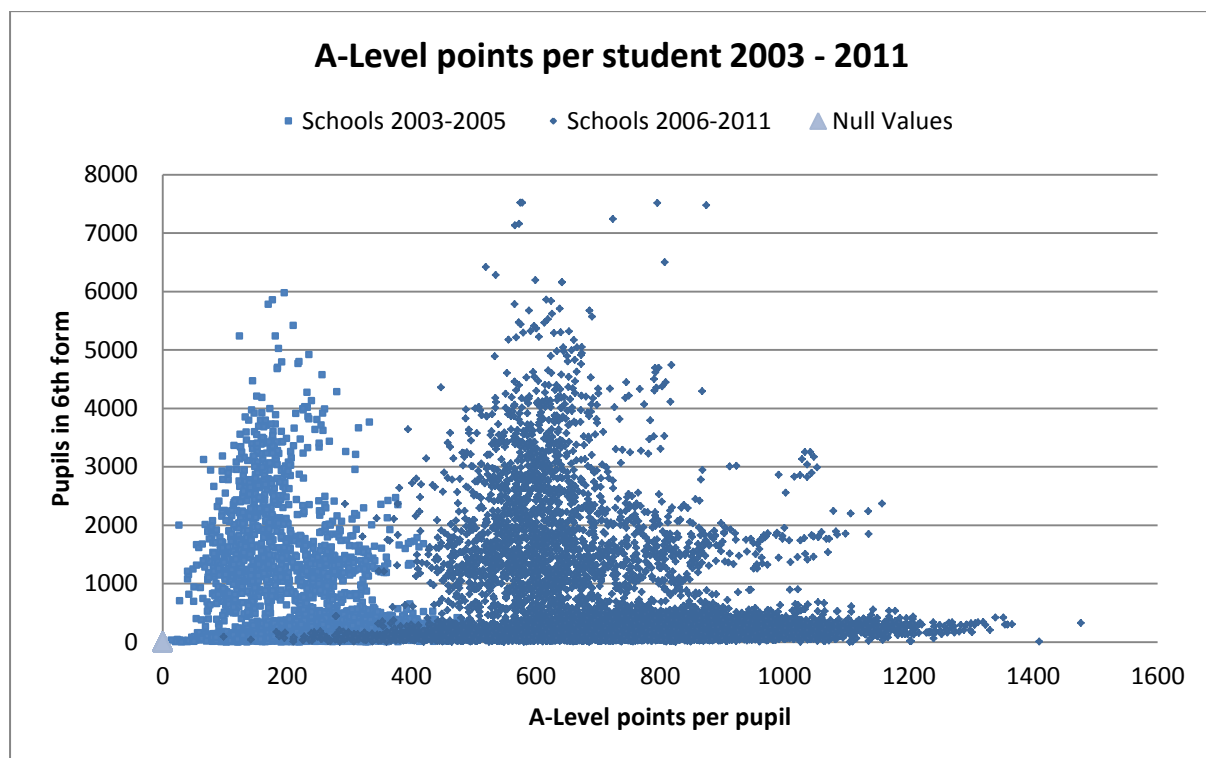


Figure 3: Scatter plot A-Level points per pupil - Pupils in 6th form 2003-2011

The two accumulations can be explained by the change of the point system from UCAS tariff to QCA tariff henceforward school year 2006 (Department for Education, 2012). To provide comparability for the further analysis, a translation between the two tariff systems is

inevitable. However, with regard to the limited time, the school datasets from 2003 to 2005 are excluded from further analysis, as the datasets over the years 2006 to 2011 comprise enough school records to offer representative analysis results.

The null values can be ascribed to missing attribute values in the records of Special schools and are excluded, as they would distort the further analysis results. Figure 4 shows the recreated sample school dataset, which comprises the school years 2006 to 2011 and is adjusted for missing values.

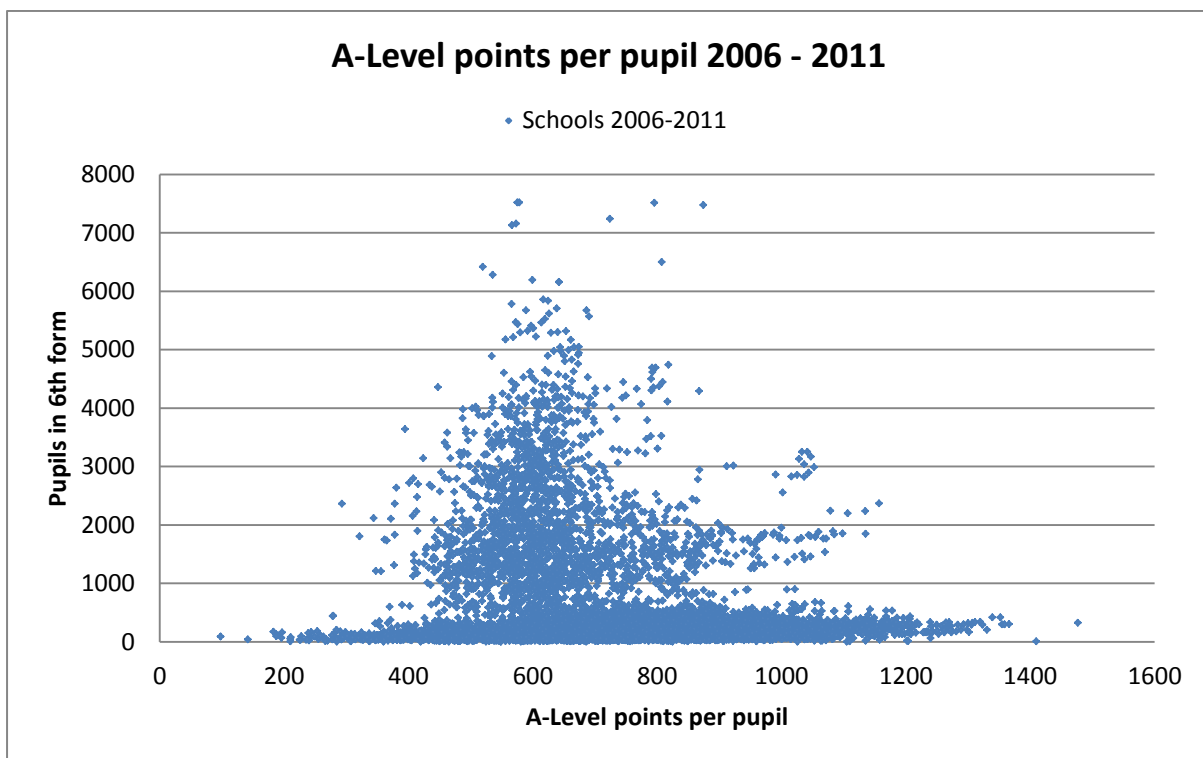


Figure 4: Scatter plot A-Level points per pupil - Pupil in 6th form 2006-2011

The objective of this stage of data analysis is the detection of apparent patterns and outliers. Except for a few cases, all schools over 1000 pupils in 6th form are General Further Education Colleges, Sixth Form Colleges or Tertiary Colleges. The negative outliers among the A-Level values are all ascribed to mixed gender Community Schools, whereas the positive outliers are ascribed to Independent Schools. A clear correlation between the number of pupils in 6th form and the performance of a school is not shown in the scatter plot.

Revealing this potential relation between the school's performance and its school type and gender policy, a breakdown analysis for the variable *A-Level points per pupil* is conducted. Therefore, descriptive statistics for the dependent variable are calculated in several data subgroups, clustered by the independent variables *Gender* and *School type*.

While clustering the data records depending on their nominal attribute *School type*, a change in coding for the year 2011 is discovered. Besides the different notation, new school types are added, making the comparability to the datasets of previous years ambiguous. Therefore the descriptive statistics are only calculated for a sample school data set that comprises the records of the years 2006 to 2010.

To begin with the breakdown analysis, the data records are first clustered by gender as well as school type. The school type clusters are then again clustered by gender. The school performance within the clusters is compared by calculating the arithmetic mean for the *A-Level points per pupil* attribute. As depicted in Figure 5, the amount of data records, assigned to the subgroups varies between the different clustering criteria, due to missing values.

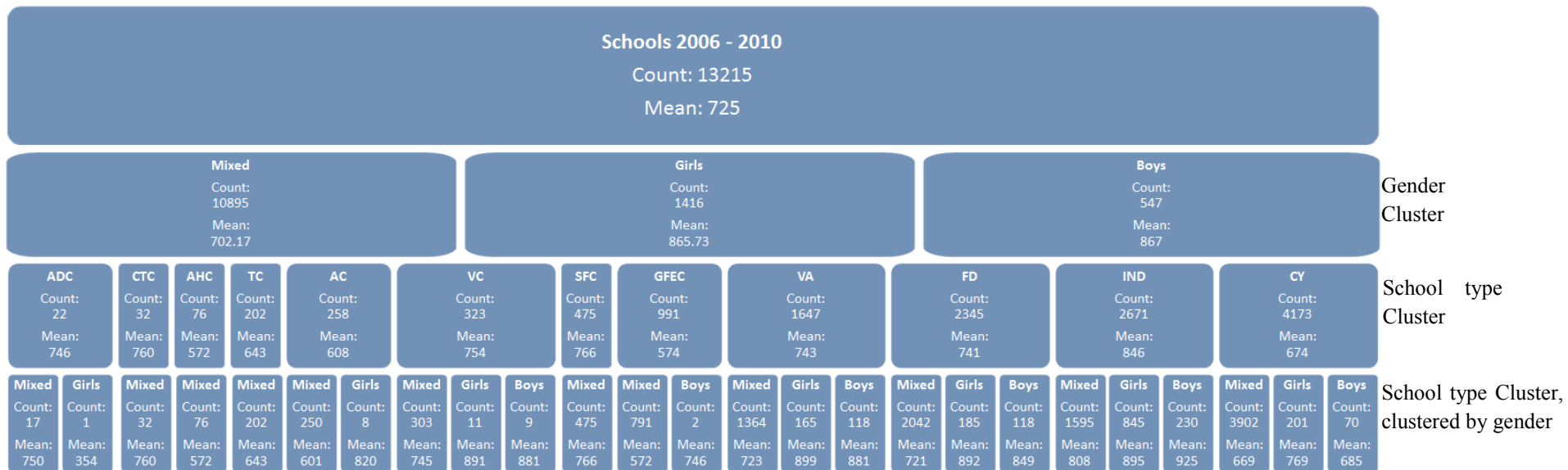


Figure 5: Breakdown analysis - amount of data records and means 2006-2010

Major findings resulting from these statistics are:

- Single-sex schools score far beyond the point average of the whole dataset and mixed gender schools, whereas there is no clear difference between boys and girls schools
- Independent schools score best, by far, whereas Agriculture and Horticulture, as well as General Further Education Colleges and Academies score far below the point average of the whole dataset
- Foundation, Voluntary aided and controlled Schools, Sixth Form Colleges and City Technology and Art and Design Colleges have about the same average, slightly beyond the point average of the whole dataset, whereas Community Schools and Tertiary Colleges score below average

- Except for Independent Schools, only girls schools perform better than only boys ones
- Single-sex Foundation and Voluntary aided and controlled Schools perform better than mixed Independent Schools
- Only girls Community Schools score much better than only boys ones, beyond the point average of all subgroups clustered by school type, except for Independent Schools

As the mean is vulnerable to outliers, further statistics are visualized, to gain insight in the distribution of school performance within the different clusters. Therefore, the quartiles of the *A-Level points per pupils* values are calculated, dividing the data set in four equal groups; the 2nd quartile is the median and separates the higher half of the data sample from the lower half. These statistics are resistant to outliers and report the dispersion of the data. As visualization technique, Boxplots that summarize these main features are used (Frigge, Hoaglin, & Iglewicz , 1989).

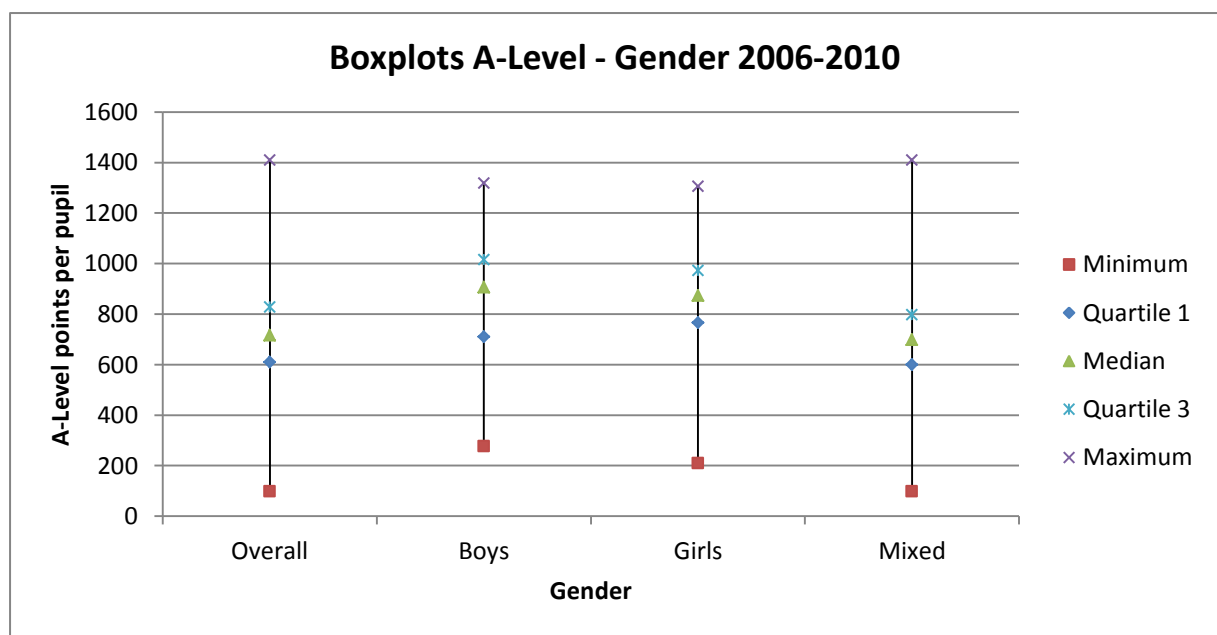


Figure 6: Boxplots A-Level points per pupil per gender 2006-2010

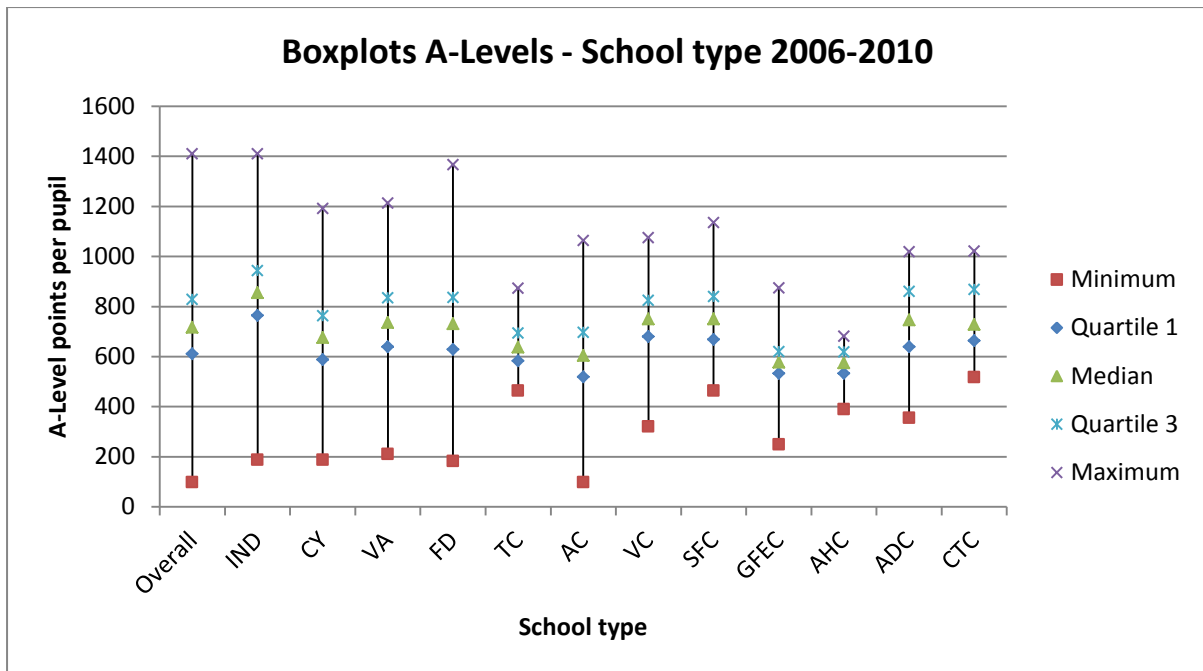


Figure 7: Boxplots A-Level points per pupil per school type 2006-2010

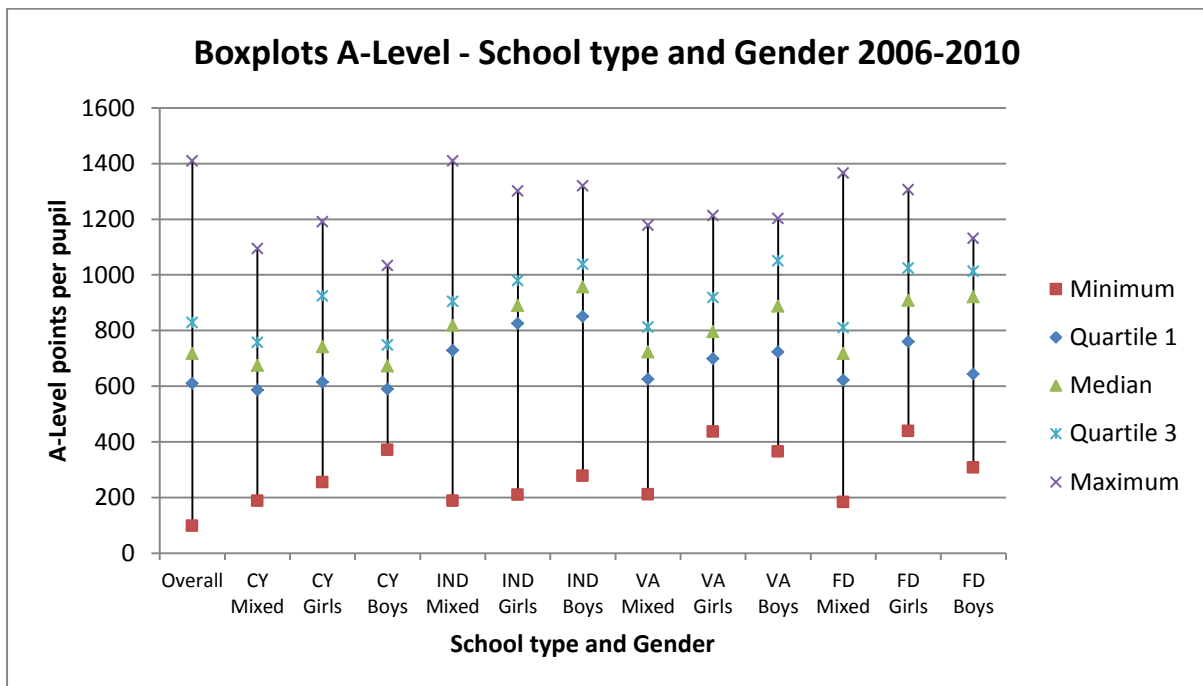


Figure 8: Boxplots A-Level points per pupil per school type and gender 2006-2010

Before those boxplots are examined in detail, the quartile dispersion coefficient as relative measure for the dispersion of the different datasets is calculated.

$$\frac{Q3 - Q1}{Q3 + Q1}$$

Equation 1: Quartile dispersion coefficient

In addition, the coefficient of variation, which is defined as the ratio of the standard deviation σ to the mean μ , is calculated.

$$c_v = \frac{\sigma}{\mu}$$

Equation 2: Variance coefficient

In contrast to the quartile dispersion coefficient, it measures the dispersion of a variable to its mean and is vulnerable to outliers (Gupta, 2009). Comparing the variance coefficient to the quartile dispersion coefficient, the influence of outliers on the dispersion can be discovered. To enable easy comparison, the coefficients of the different subgroups within the established clusters are visualized in the bar charts below, starting with the gender cluster.

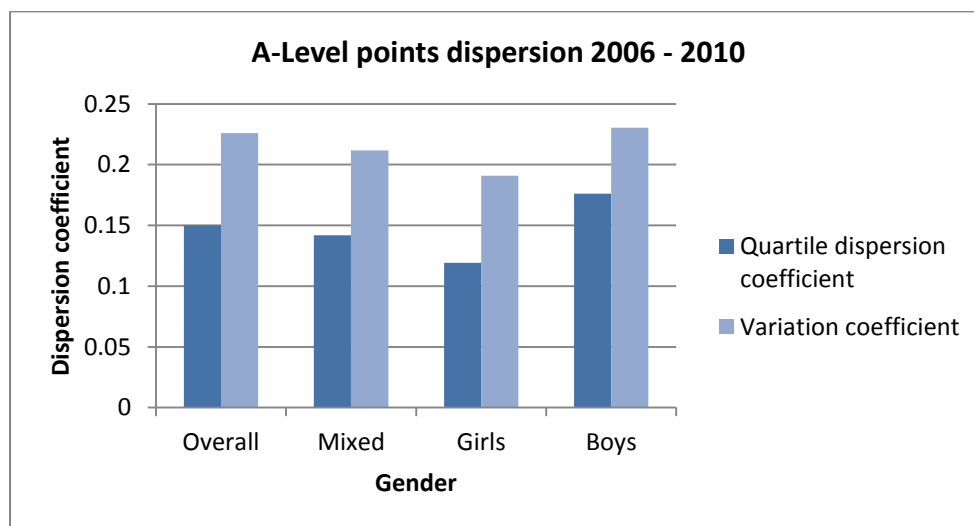


Figure 9: Dispersion coefficients A-Level points per pupil per gender 2006-2010

The dispersion of the performance values of boy schools is higher than of the whole dataset, comparing the quartile dispersion coefficients. As the difference between the quartile dispersion and variation coefficient of boy schools is low, the relatively high dispersion is not caused by outliers. Looking at the boys school boxplot in Figure 6, this dispersion is caused by a relatively high dispersion in the second quartile. Regarding the low quartile dispersion coefficient, the performance of girl schools is quite concentrated around the medium. This conclusion is confirmed by analysing the according boxplot.

Following the quartiles and dispersion coefficients for the subgroups clustered by school type are visualized.

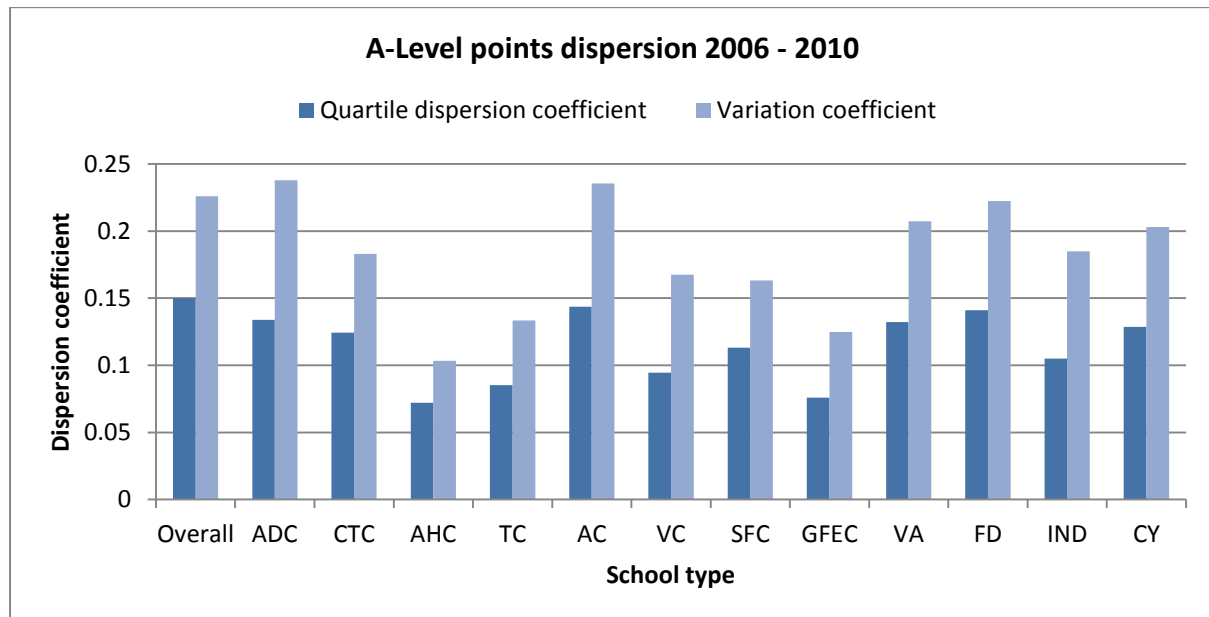


Figure 10: Dispersion coefficients A-Level points per pupil per school type 2006-2010

All quartile dispersion coefficients of the different subgroups, clustered by school type, are lower than the coefficient of the whole dataset. For Art and Design, Academies, Voluntary aided, Foundation and Community Schools, the dispersion of the performance values from the mean is quite high. The high difference between the quartile dispersion and variation coefficient of the performance values of those school types as well as Independent and Voluntary controlled Schools is a sign of outliers in those subgroups. The dispersion of performance values of Tertiary Colleges, Agricultural and Horticulture and General Further Education Colleges is relatively low. Those assumptions are mainly approved, regarding the boxplots in Figure 7. However, Voluntary controlled schools as well as Art and Design Colleges do not have extreme outliers in comparison to Independent, Community, Voluntary aided and Foundation Schools.

To finalize the breakdown analysis, the descriptive statistics are also calculated and visualized for the subgroups clustered by school type and gender. A comparison is only practicable for Community, Foundation, Independent and Voluntary aided Schools, as those are the only school types with single-sex policy, among the data records.

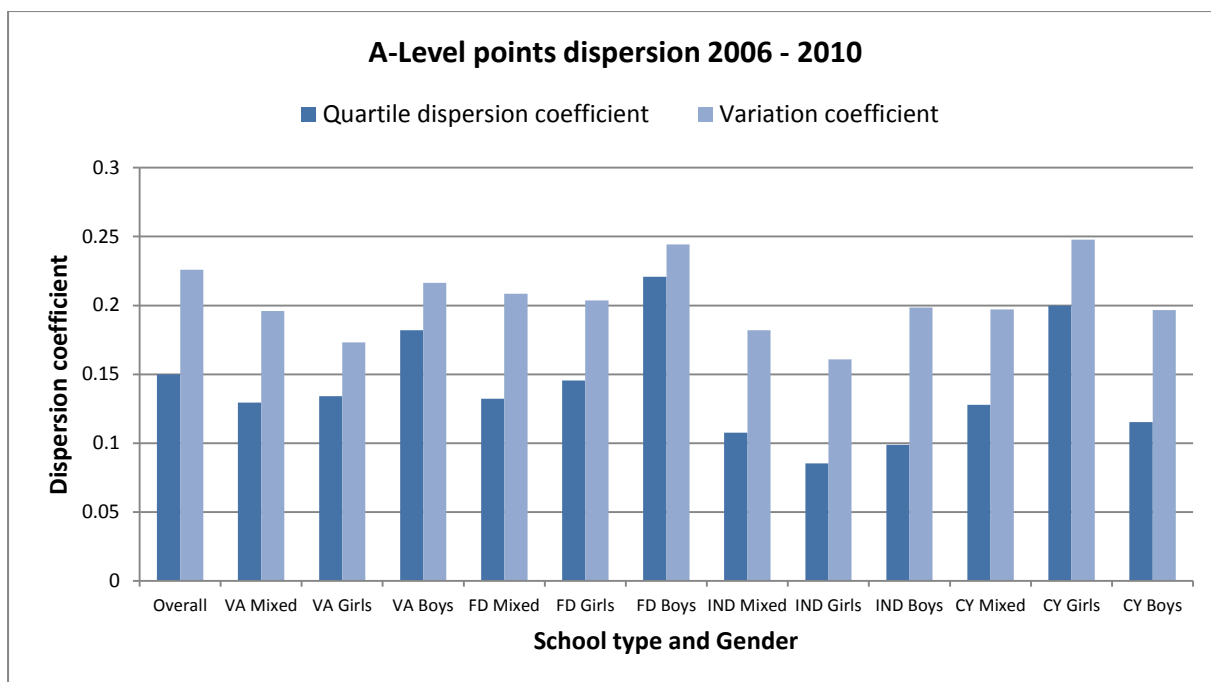


Figure 11: Dispersion coefficients A-Level points per pupil per school type - gender

In comparison to other school types, the dispersion of performance values of Independent Schools is low, except for clear outliers. Especially for Foundation and Voluntary aided boys Schools as well as Community girls Schools the quartile dispersion is high. These observations are also approved by the visualization of the data in the boxplots in Figure 8. Apart from Independent Schools, the dispersion within the school types clustered by gender is different than the dispersion in the gender subgroups in Figure 9. The quartile dispersion coefficient of mixed schools is lower than of girls schools, and Community Schools also show higher dispersion among girls schools than boys schools. With exception of girls Community Schools, the ratio of the variation coefficients is the same as in Figure 9. Comparing the dispersion among the same gender of the different school types with the dispersion of the subgroups, clustered only by school type in Figure 10, the ratio of the coefficients is consistent except for single-sex Community Schools. This observation leads to the presumption, that A-Level performance of a school might be more correlated to its school type than gender, as the dispersion of the *A-Level points per pupil* within the combined cluster is more similar to the school type cluster.

Regional data

To start with the analysis of the regional data, the UK Competitive Index of the years 2005, 2006 and 2008 to 2010 is visualized in a geographic map in Google Fusion Tables. As the UKCI is specified for local authority areas, a map of England is overlaid with the

corresponding borders, using a public .kml⁵ file. The UK Competitive Index tables are imported in Google Fusion tables and merged with the .kml file, using the *Local authority area* attribute as join. While merging the tables, minor changes of the allocation of local authority areas henceforward 2009 are detected. As this concerns only few areas, these are excluded from further analysis. To visualize the UKCI on the map, the bordered areas are highlighted according to the UKCI quartiles they belong to. This is realized by applying Google Fusion Tables' customized filter function for polygons.

UKCI	Purple	Orange	Yellow	Green
2005	<=90	90<=97	97<=107	107<
2006	<=90	90<=97	97<=106	106<
2008	<=89	89<=97	97<=106	106<
2009	<=89	89<=97	97<=106	106<
2010	<=89	89<=97	97<=105	105<

Table 4: Quartiles UKCI 2005-2010

The excluded, incomparable areas for the years 2005, 2006 and 2008 are highlighted in red.

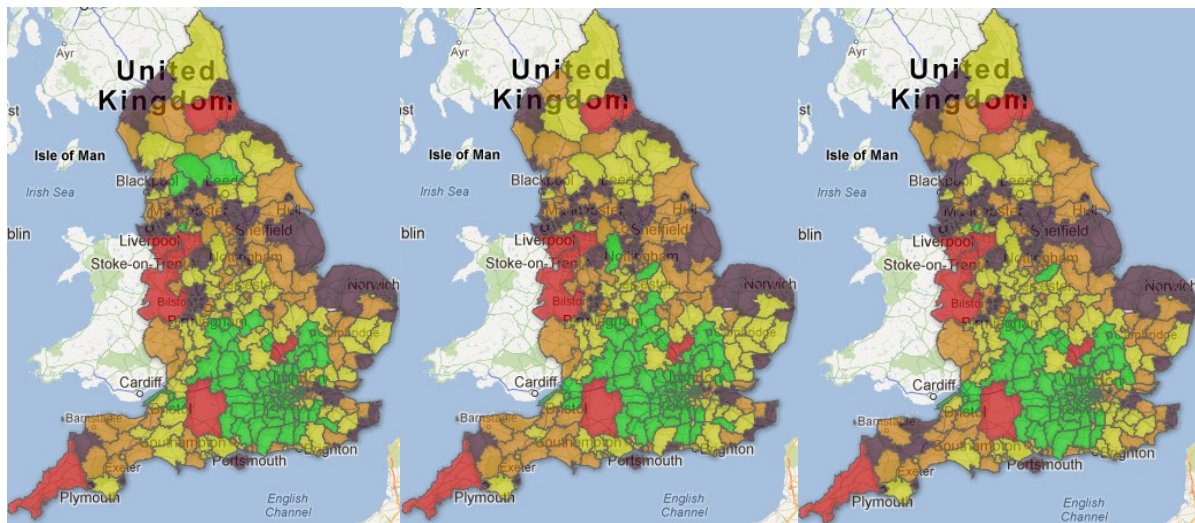


Figure 12: UKCI geographic map 2005, 2006 and 2008

⁵ Keyhole Mark-up Language: originally developed for geographic visualization in Google Earth

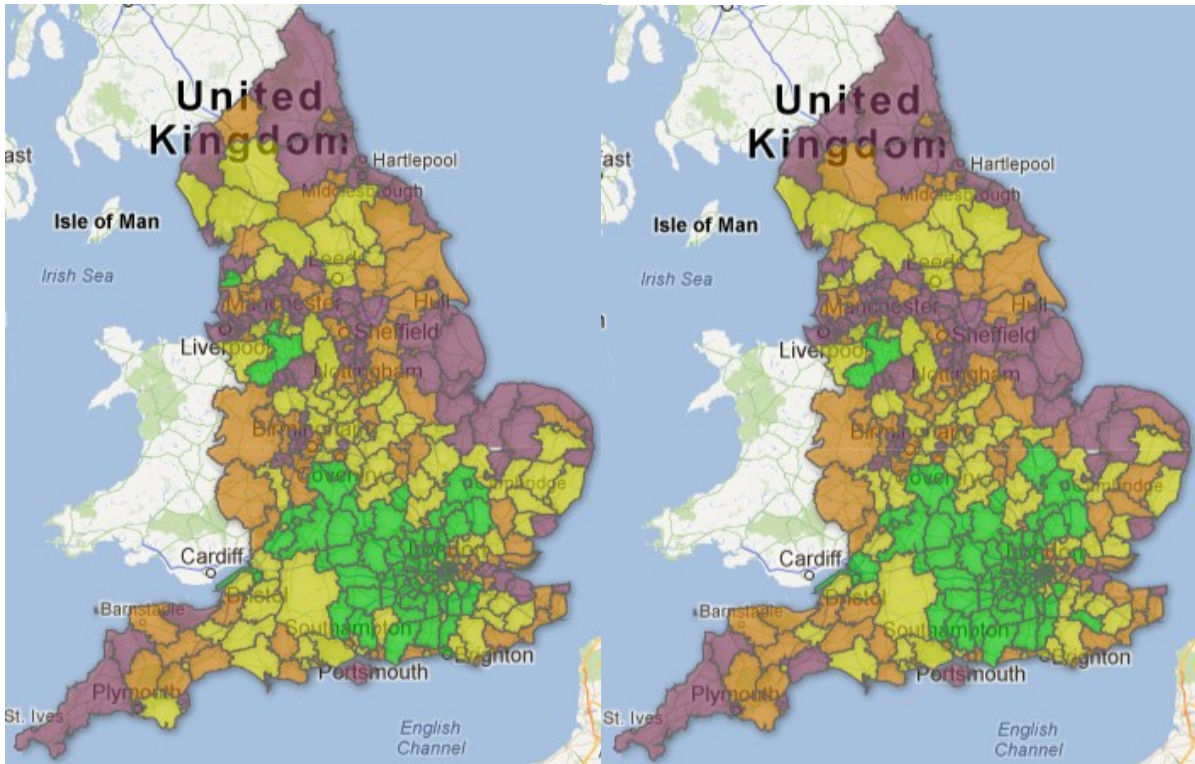


Figure 13: UKCI geographic map 2009 and 2010

At first glance, the Competitiveness Index of the different regions remains constant over the years. The most competitive areas are London and its urban hinterland, as well as the joining western, northern and southern districts. The less competitive districts are located in South and North England, at the east coast and around Manchester and Sheffield.

Having gained a first understanding of the UKCI, it is visualized in a scatter plot to detect outliers and null values.

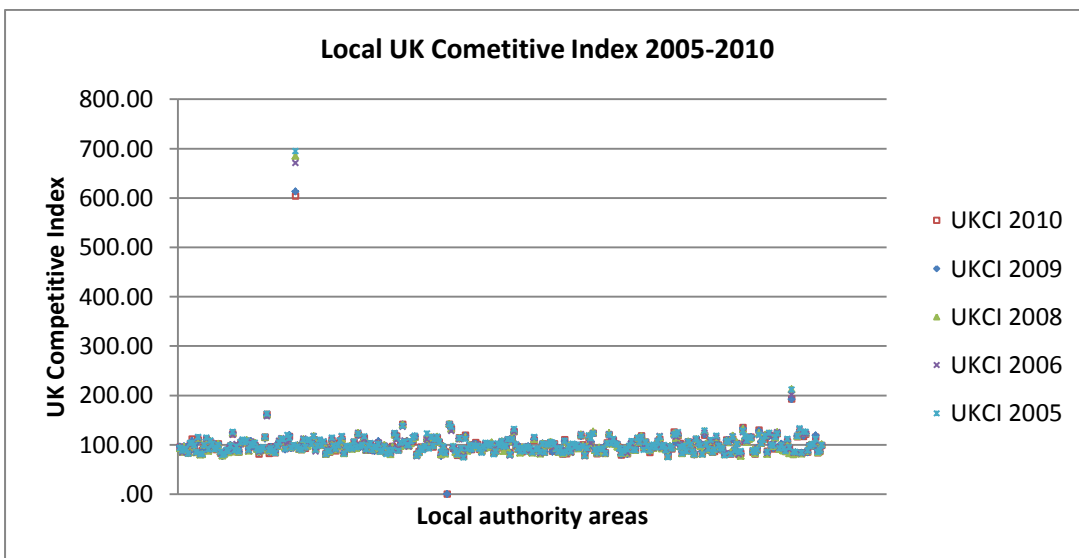


Figure 14: UKCI scatter plot 2005-2010

The extremely high outlier is assigned to the City of London; the null value is caused by a missing value assigned to the Isle of Sicily.

School data in geographic context

After the school and regional datasets were visualized separately, a first step towards their integration is done, displaying the location of schools on a geographic map. As sample, each school of the school dataset 2010 is mapped by a pin on the map below.⁶ The performance of the school is distinguished by the colours of the pins, corresponding to the quartiles of the *A-Level points per pupil* attribute of the 2010 dataset.

A-Level points per pupil	Red	Pink	Yellow	Green
Quartiles	≤ 635	$635 < \leq 733$	$733 < \leq 841$	$841 <$

Table 5: Quartiles A-Level points per pupil 2010

For comparison, the map is displayed next to the map which displays the distribution of the UKCI 2010.⁷

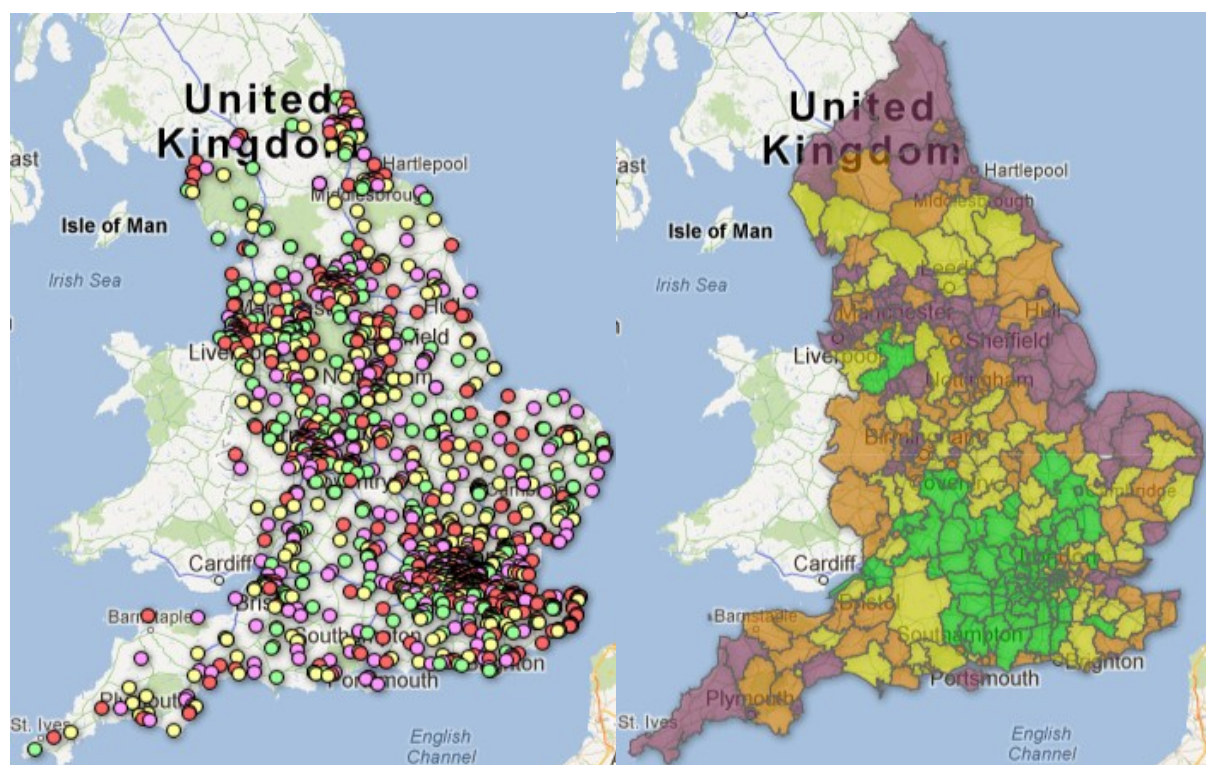


Figure 15: Geographical comparison A-Level points per pupil – UKCI 2010

⁶ The maps of the years 2006, 2008 and 2009 are in the appendix

⁷ An integrated visualization in Google Fusion Tables is only feasible by publishing the tables in the web

At a first glance, no clear patterns within the geographical distribution of schools with similar performance are identified, nor are correlations between the performance of local schools and the local UKCI.

This visualization of the different datasets provided an insight in the data records and disclosed interesting potential research questions that could be further investigated upon. However, before the research questions for this project are specified in detail, the data quality is examined in the next chapter.

3.1.3. Data Quality

The quality evaluation of the datasets is based on the quality components defined by Eurostat (Bergdahl, et al., 2007) and Wang, Storey and Firth (1995):

1. COMPLETENESS of the dataset and its attributes
2. RELEVANCE corresponding to research question
3. ACCURACY according to true values
4. COMPARABILITY over time, domains and between geographical areas
5. COHERENCE of data combined from different sources
6. PHYSICAL ACCESSIBILITY and clarity provided by meta data

Based on those criteria, the quality of the A-Level performance tables and UKCI dataset is analysed in the table below.

Quality criteria	A-Level performance tables	UKCI dataset
Completeness	Datasets do not comprise all English schools Missing values for <i>A-Level points per pupil</i> for Special schools Scattered missing values for other attributes in the dataset	No dataset for year 2007 Missing values for Isle of Sicily
Relevance	High, offers school profile, including performance indicators	High, UKCI as indicator for socio-economic standing of an area
Accuracy	Accuracy is limited by statistical errors	Accuracy of the UKCI components is limited by statistical errors UKCI measured by simplified three factor model, entire complexity of socio-economic interdependencies is not comprised
Comparability	Comparability over time hindered due to changing A-Level point tariffs between 2005 and 2006	Comparability of some areas over time hindered due to changing local authority areas

	Different values for <i>School type</i> attribute in the 2011 dataset	
Coherence	Data from Department for Education	UKCI components from Office for National Statistics
Accessibility and clarity	Good accessibility, clear meta data is provided	Good accessibility, clear meta data is provided

Table 6: Quality assessment A-Level performance tables and UKCI dataset

In summary, the quality of both dataset is on a level that facilitates an informative, valid and significant data analysis. Subsequently, the research question is specified, closing the project stage of data understanding.

3.1.4. Research Question Detailing

The visualization of the data, relevant for the preliminary research topic selected in Chapter 3.1.1 - Factors influencing a school’s performance in A-Levels - already revealed interesting patterns and potential correlations between the data attributes. As the quality of the A-Level performance tables as well as the UKCI datasets is considered adequate, the research topic is now specified in detail. The term “factors” is particularized in the terms “school properties” – implying the attributes of the performance tables – as well as “UK Competitive Index” – representing the regional factors. Subsequently, two research questions are defined:

1. Analysis of correlations between A-Level performance of schools and school properties
2. Analysis of the correlation between A-Level performance of schools and the UK Competitive Index

In the next process step, good quality data and data attributes, which are relevant to the research questions, are selected from the available datasets, based on the findings from the data understanding stage in Chapter 3.1.

3.2. Data Selection

The school data relevant to the research question is comprised in the A-Level performance tables. Due to the different A-Level point tariff system in the datasets 2003 to 2005 and the incomparability of the *School type* attribute in the dataset 2011, the sample dataset for this analysis is composed of the datasets 2006 to 2010. In addition, to investigate on the second research question, all accessible datasets for the UK Competitive Indicator are selected, comprising the years 2005, 2006 and 2008 to 2010. The data attributes, regarded as relevant

for the further data analysis, are summarized in Table 7. The selection is made, based on the completeness, significance and relevance of the attributes for the research questions.

Attribute	Selection criteria	Data type	Dataset
Year	Required for correlation analysis with UKCI	Ordinal	A-Level performance tables 2006-2010
School_Postcode	Foreign key to join A-Level performance and UKCI tables	Nominal	A-Level performance tables 2006-2010
School_Type	Potential relation to A-Level performance	Nominal	A-Level performance tables 2006-2010
School_Admission_Policy	Potential relation to A-Level performance	Nominal	A-Level performance tables 2006-2010
School_6thForm_Gender	Potential relation to A-Level performance	Nominal	A-Level performance tables 2006-2010
Pupils_in_6 th _Form	Potential relation to A-Level performance	Numeric	A-Level performance tables 2006-2010
Alevel_Points_per_Pupil	A-Level performance indicator of a school	Numeric	A-Level performance tables 2006-2010
Local Authority Area	Foreign key to join A-Level performance and UKCI tables	Nominal	UKCI tables 2005, 2006, 2008-2010
UKCI	Socio-economic performance indicator of a region	Numeric	UKCI tables 2005, 2006, 2008-2010

Table 7: Data selection

In the next process step, a data analysis model, which investigates the correlations between A-Level performance and school properties, and the UKCI respectively, is developed.

3.3. Data Analysis Model Development

The investigation on the dependencies between school performance and different school attributes as well as the UKCI requires a correlation analysis. In contrast to the school properties, the impact of the socio-economic factors on the performance of a school is likely to be time-delayed. Therefore, two models are developed, each adapted to one of the two research questions.

3.3.1. Correlation between A-Level performance and school properties

The development of a data analysis model starts with the selection of the model class that determines the structure of the data analysis result (Berthold, Borgelt, Hoepfner, & Klawoon, 2010). In statistics, the degree of correlation between two numerical attributes is measured by correlation coefficients. Different coefficients, which are sensitive to different mathematical functions, exist. Linear dependencies between two attributes are measured using the Pearson correlation coefficient, whereas Spearman's rank coefficient gives evidence of a monotonic

function between two quantitative attributes. In contrast to correlation coefficients, the correlation ratio measures the correlation between nominal and quantitative attributes (Mirkin, 2011). The table below indicates the methods used for the correlation analysis between A-Level performance and school properties.

Attributes	A-Level points per pupil
Pupil in 6 th form	Spearman's rank coefficient and Pearson correlation coefficient
School type	Correlation ratio
Gender	Correlation ratio
Admission policy	Correlation ratio

Table 8: Correlation analysis methods

The relationship between the numeric attributes *Pupils in 6th form* and *A-Level points per pupil* is investigated using both, Spearman's rank and the Pearson correlation coefficient. For a paired sample dataset (X_i, Y_i) of size n the Sample Pearson correlation coefficient is defined as (Rodgers & Nicewander, 1988):

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Equation 3: Sample Pearson correlation coefficient

Spearman's rank coefficient that testifies if increasing values for one attribute lead to increasing values of the other and accordingly decreasing values for one attribute lead to decreasing values of the other is defined as the Pearson correlation coefficient but between ranked pairs of data (x_i, y_i) (Park & Lee, 2001):

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Equation 4: Spearman's correlation coefficient

These correlation coefficients are calculated for the *Pupils in 6th form* and *A-Level points per pupil* values of the whole dataset, as well as for each subgroup of the breakdown analysis of Chapter 3.1.2 - enhanced by the attribute *Admission policy* - to reveal the dependencies for different school types, gender and admission policies. The coefficients take on a value between -1 and 1, in which the sign points to the direction of the relationship between the investigated attributes; the higher the absolute value, the higher the correlation between the paired data.

The correlation between *A-Level points per pupil* and the categorical attributes *Gender*, *School type* and *Admission policy* is measured, using the ratio of the attribute's average variance of the categorical group - σ_w^2 - and its variance within the whole dataset - σ^2 (Mirkin, 2011):

$$\eta^2 = 1 - \frac{\sigma_w^2}{\sigma^2}$$

Equation 5: Correlation ratio

This correlation ratio is applied on the *A-Level points per pupil* variances of each cluster, established in the breakdown analysis, to reveal and compare potential dependencies between school performance and different school properties. The ratio takes on a value between 0 and 1 and states the percentage of the *A-Level points per pupil* value that is determined by the independent variables. The square root of the correlation ratio informs about the correlation (Mirkin, 2011).

3.3.2. Correlation between A-Level performance and UK Competitive Index

Before the correlation between A-Level performance of schools and the UK Competitive Index is analysed in particular, the correlation between performance and location of a school is calculated, using the correlation ratio as stated in Equation 5. Thereby, the average variance of *A-Level points per pupil* of each local authority area is divided by the variance within the whole dataset.

Thereafter, an analysis model investigating the correlation between A-Level performance of schools and the UK Competitive Index is established. As there might be a time gap between the dynamics and subsequent impacts of the attributes, as well as a time delay until those are reflected in the statistical figures, a time series analysis is conducted. Therefore Spearman's Rank and the Pearson correlation coefficient - as stated in Equation 3 and Equation 4 –are calculated for the following pairs of data attributes:

- Average of *A-Level points per pupil* median over the years of each region / Average UKCI over the years of each region
- *A-Level points per pupil* median of each region / UKCI of each region

For the latter, the coefficient is calculated for each pair of datasets of the different years, to take a potential time gap into account. Furthermore, the relation between the school's A-

Level performance and the UKCI is calculated within subgroups clustered by school type and geographical entity. The school type cluster criterion is chosen, as the data visualization in Chapter 3.1.2 revealed significant differences between the performances of different school types. The clustering by geographical entities is considered, as the geographical visualization of the UKCI during the data visualization stage revealed patterns of urban superiority.

3.4. Data preparation

Before the data analysis models are deployed, the data sample that was selected in Chapter 3.2 is prepared, to optimize its quality and applicability. For the data visualization stage, the A-Level performance tables for each year are already merged to one dataset that undergoes the cleaning process, documented in the next section. The UKCI tables of the different years are merged in the course of the subsequent data integration.

3.4.1. Data Cleaning

First the A-Level data is cleaned from missing values and outliers that were partly identified during the data visualization stage. All records with missing values for the *A-Level points per pupil* attribute are excluded from the sample data set, as this attribute is essential for both analysis models. The remaining amount of records for the school type *Special school* is not representative for further analysis and excluded as well.

Beside this comprehensive dataset, several sub datasets for the breakdown analysis are prepared, excluding data records with missing values for the clustering criteria. The figure below illustrates the subgroups of the breakdown analysis with the absolute amount of complete data records.

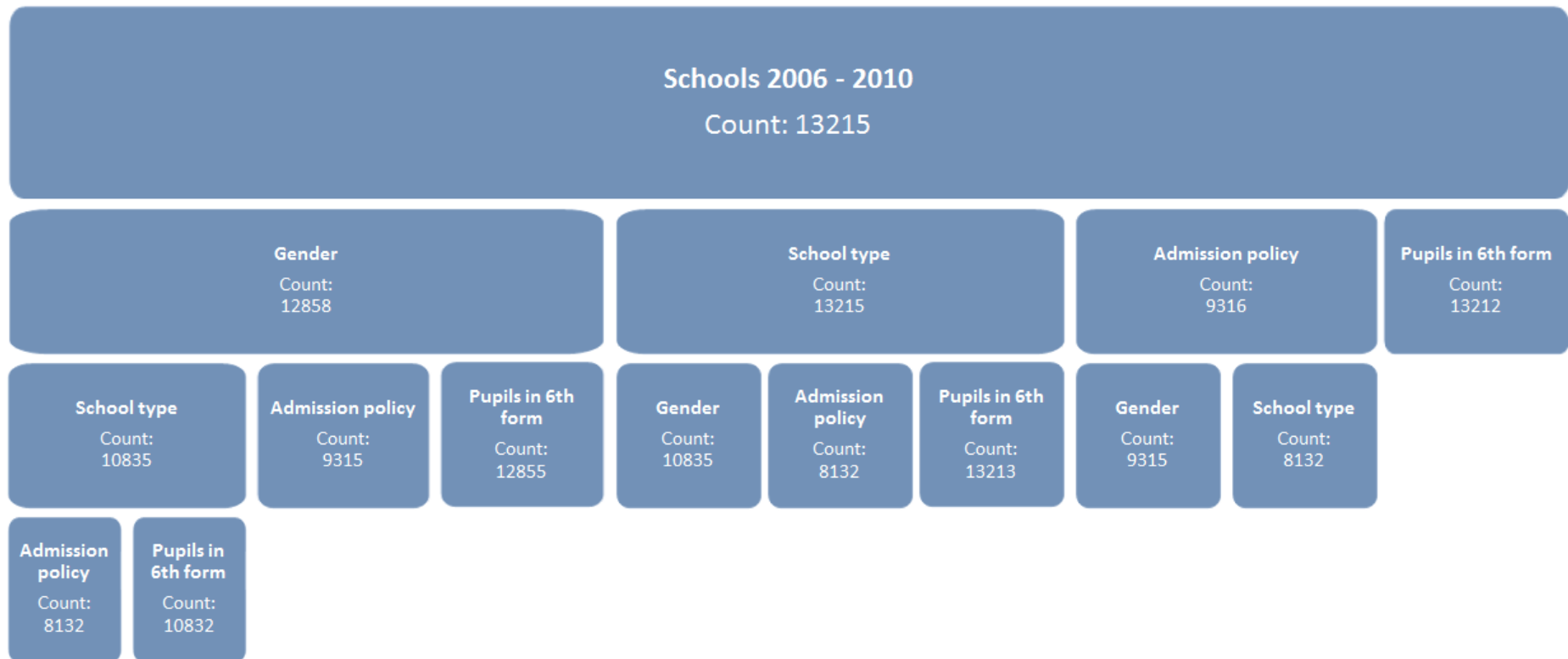


Figure 16: Data records per cluster of the breakdown analysis

The value distribution of the independent nominal data attributes within the clusters is shown in the charts below. This information is crucial to assume the representativeness and significance of the data analysis results.

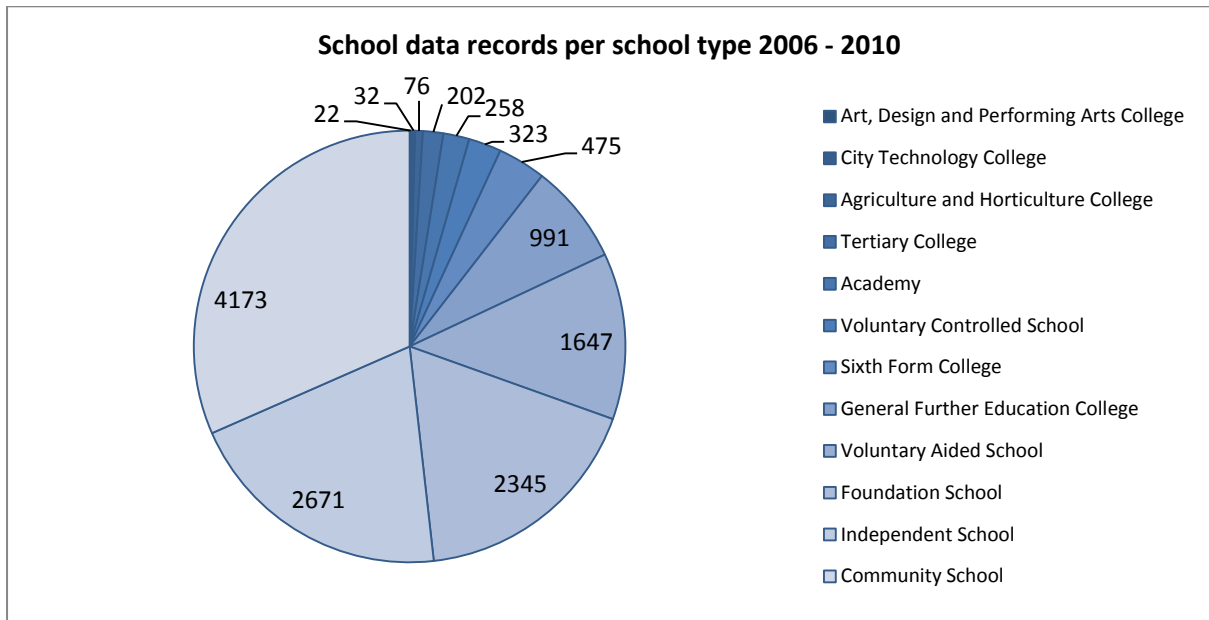


Figure 17: School data records per school type 2006-2010

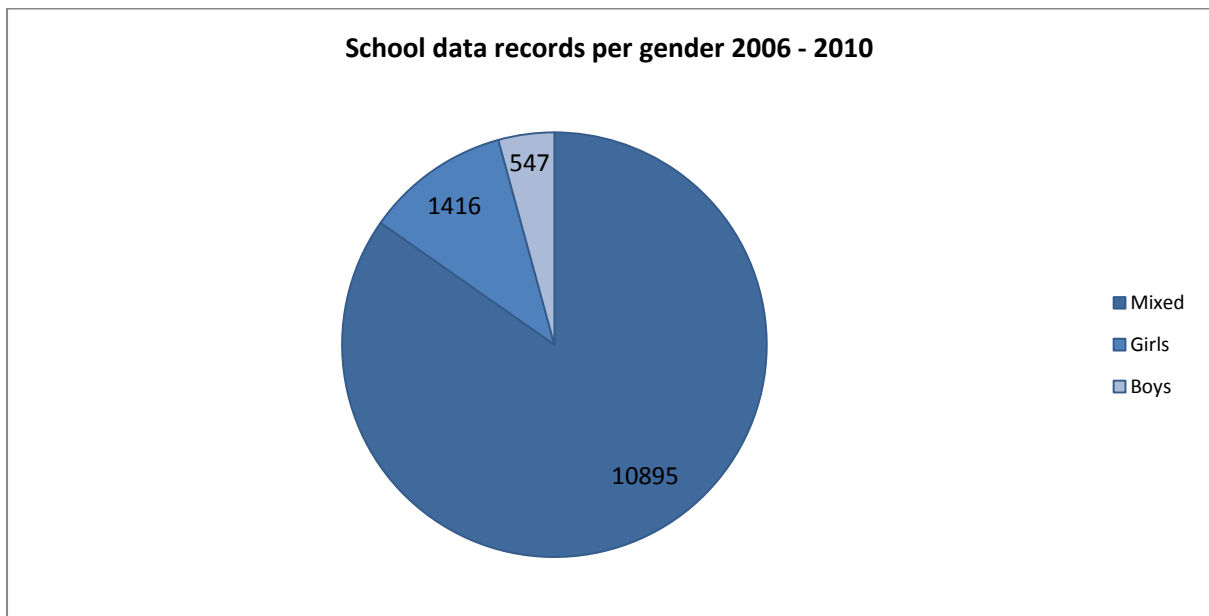


Figure 18: School data records per gender 2006-2010

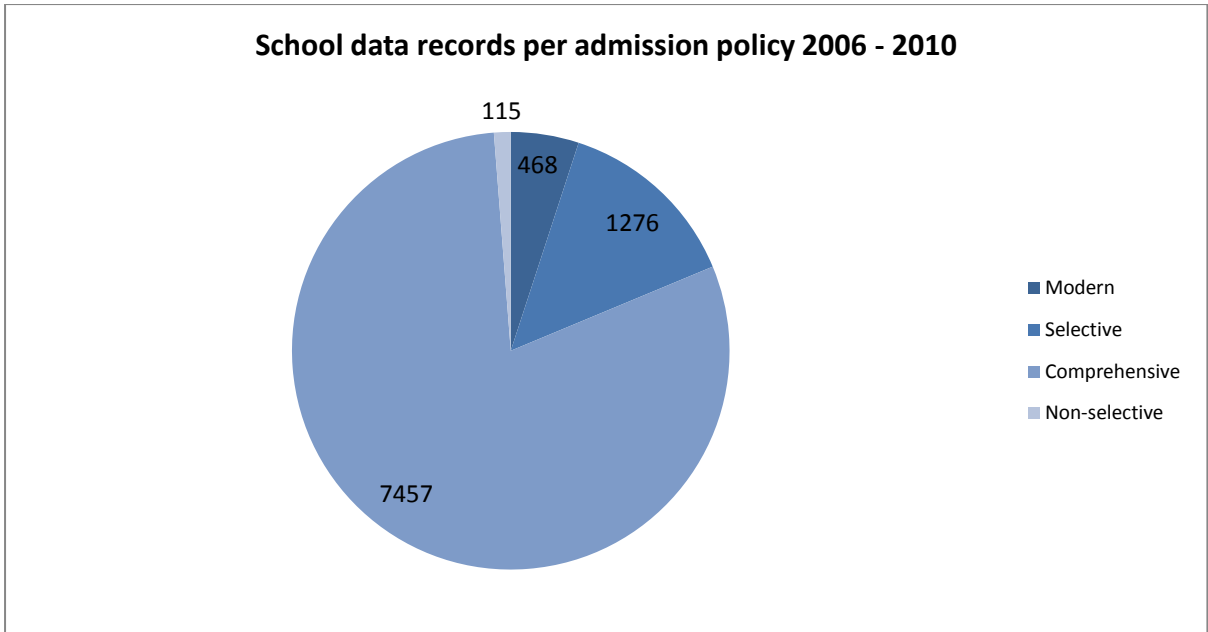


Figure 19: School data records per admission policy 2006-2010

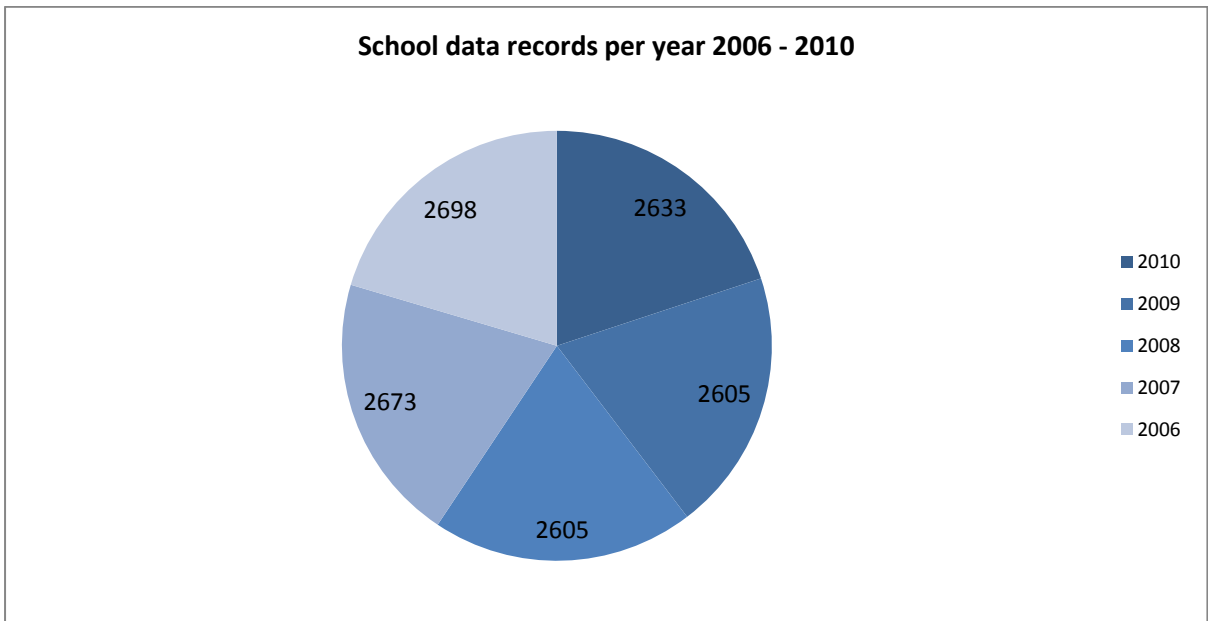


Figure 20: School data records per year 2006-2010

For the subgroup clustered by gender and school type, an appropriate amount of data records is only available for Voluntary aided, Foundation, Independent and Community Schools.

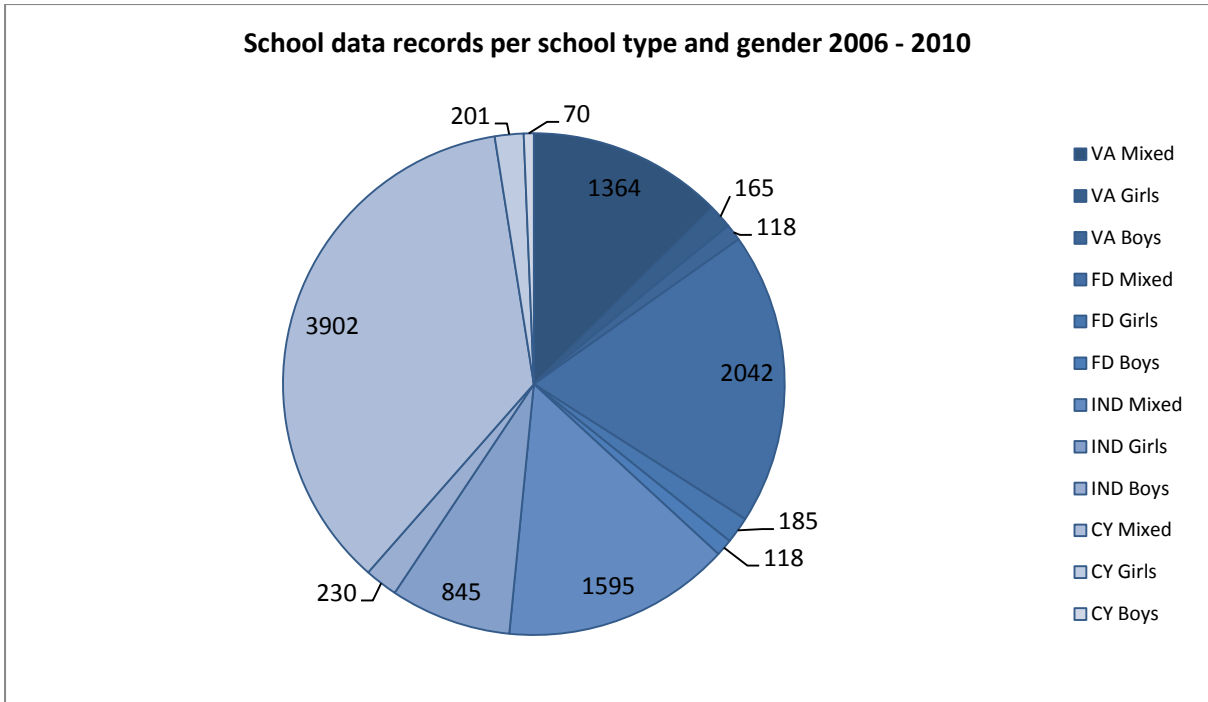


Figure 21: School data records per school type and gender 2006-2010

The clustering by school admission policy and school types is only reasonable for Community, Foundation and Value aided Schools. For the other school types, either an appropriate amount of data records is not available, or the values for the *Admission policy* attribute are uniform.

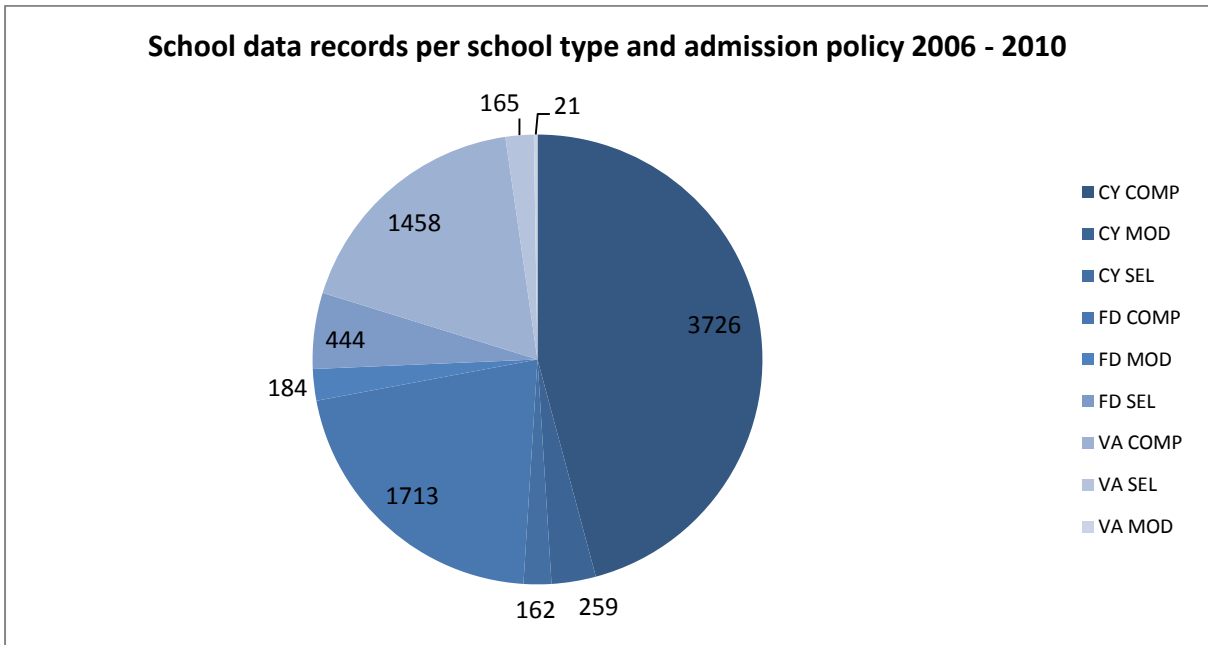


Figure 22: School data records per school type and admission policy 2006-2010

The clustering by school admission policy and gender is not conducted for non-selective admission, as not enough data for the different subgroups clustered by gender is provided.

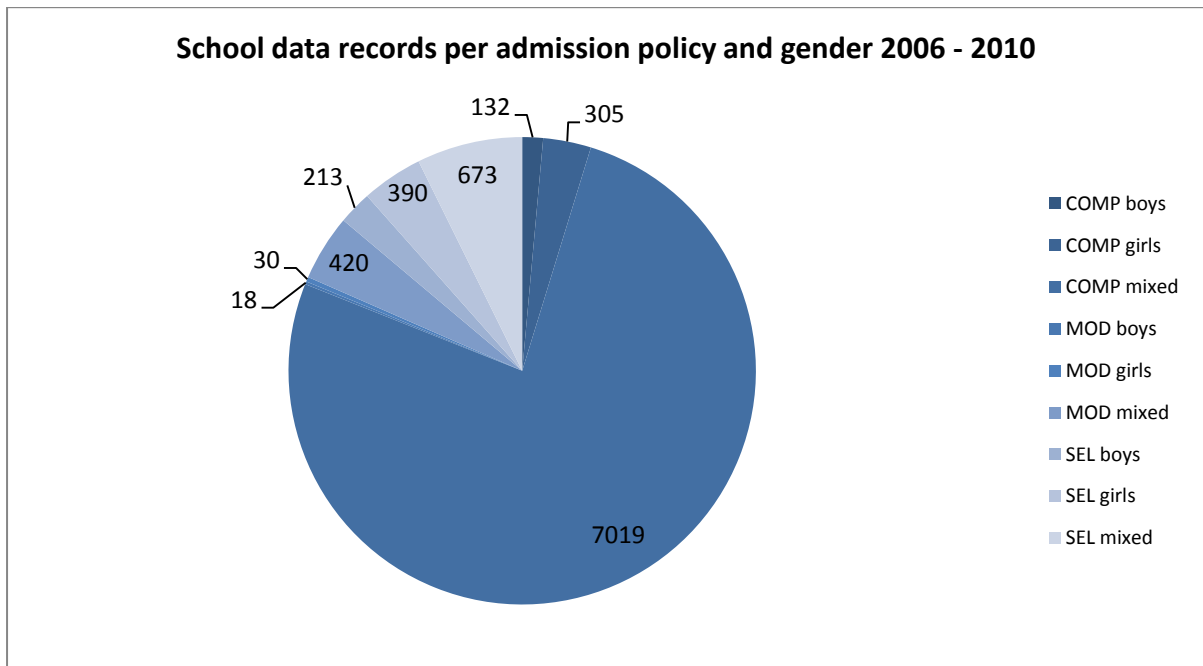


Figure 23: School data records per gender and admission policy 2006-2010

The clustering by school type, admission policy and gender is only conducted for selective and comprehensive admission policy, as an appropriate amount of data records is not available for modern admission policy.

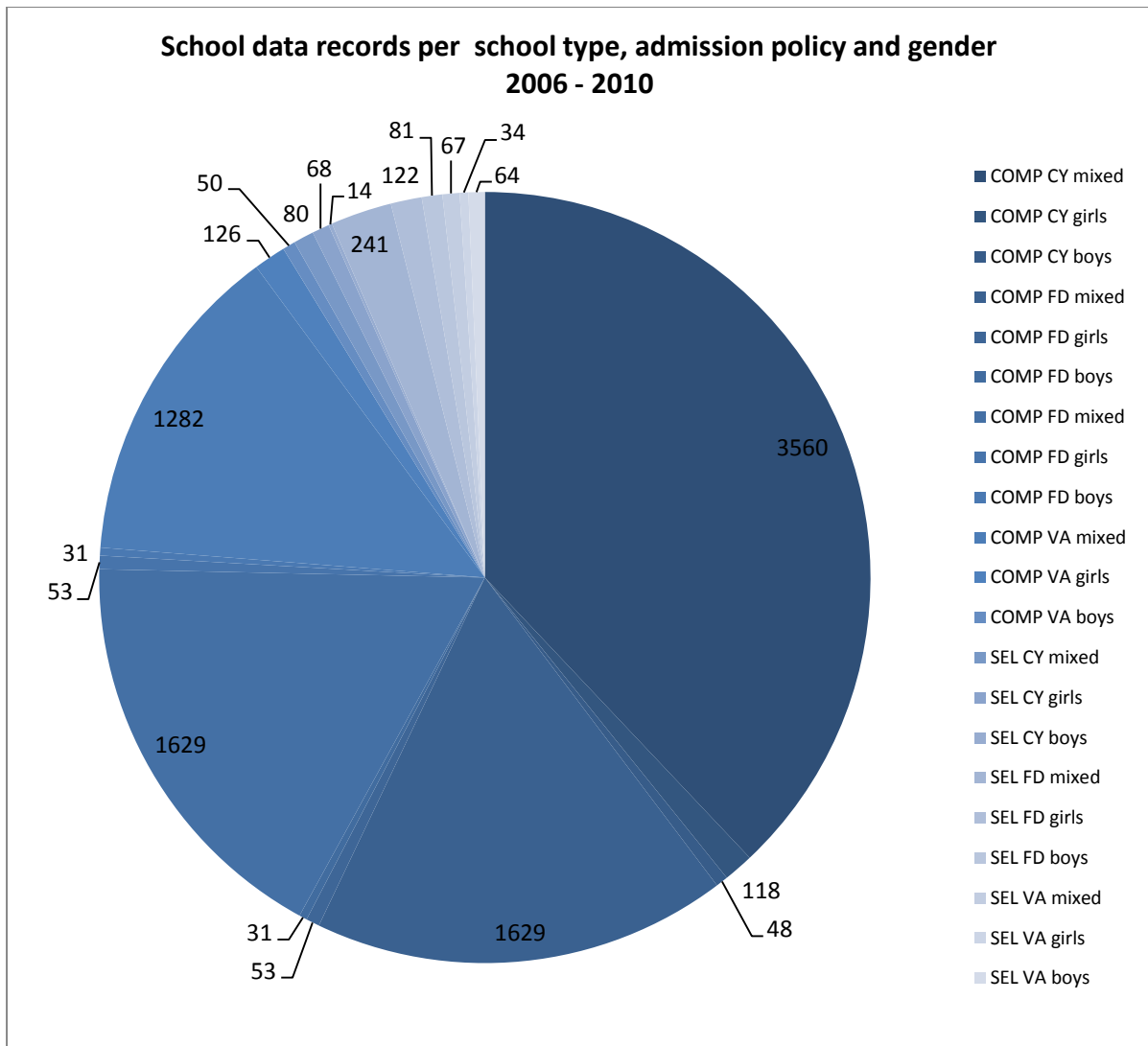


Figure 24: School data records per school type, gender and admission policy 2006-2010

3.4.2. Data Integration

The sub dataset for the correlation analysis between A-Level performance of schools and the UKCI is constructed, joining the A-Level school league and UKCI tables. First the UKCI tables of the different years are merged. As the local authority areas were subject to minor changes between 2008 and 2009, incomparable regions are deleted from the UKCI dataset. Next the UKCI and A-Level performance tables are extended by a data field for the local authority code, serving as primary key and foreign key, respectively. For this purpose a local authority code - postcode translation table, included in the Good Schools Guide metadata collection, is joined with the A-Level dataset on the *School postcode* attribute, using QlikView. To indicate the A-Level performance of each local authority areas, the median of the A-Level performance of all local schools is calculated for the data set. The medians are then horizontally integrated with the extended UKCI table, using the local authority code as

foreign key. Additionally, this horizontal integration is conducted for each subgroup clustered by school type. Regions that have missing values for the *A-Level points per pupil* attribute are deleted from the datasets. Furthermore the City of London, which was detected as an outlier in the data visualization stage, is excluded as it would distort the further analysis. Figure 25 indicates the number of A-Level performance records, which are calculated in the medians of the different school type data sets of the years 2006 to 2010 excluding 2007, compared to the general data set.

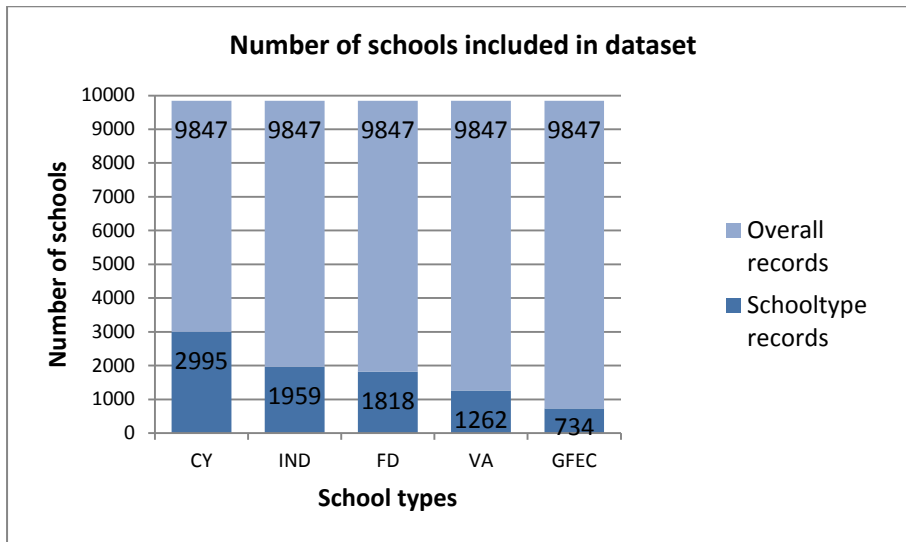


Figure 25: Number of schools included in integrated UKCI datasets

The figures below illustrate the number of local authority areas included in the different school type data sets, compared to the general data set. In addition, the missing regions are highlighted red in the attached maps.

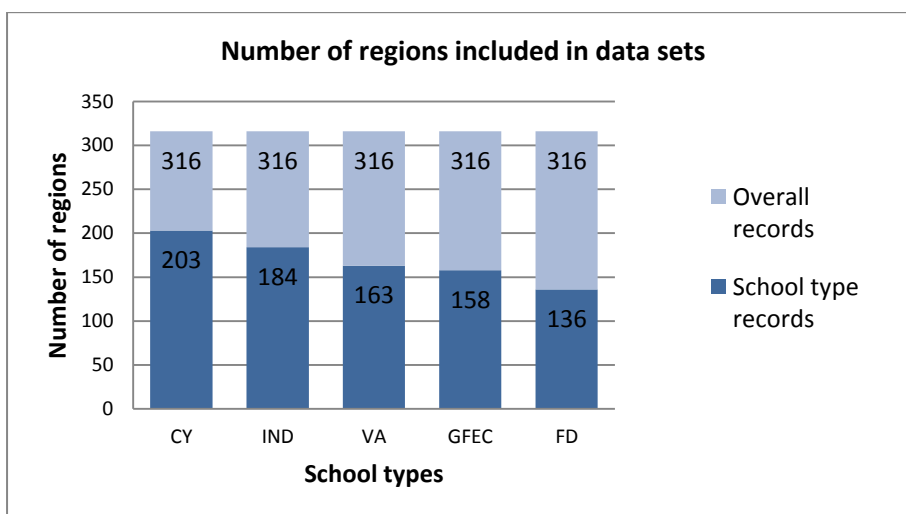


Figure 26: Number of local authority areas included in datasets

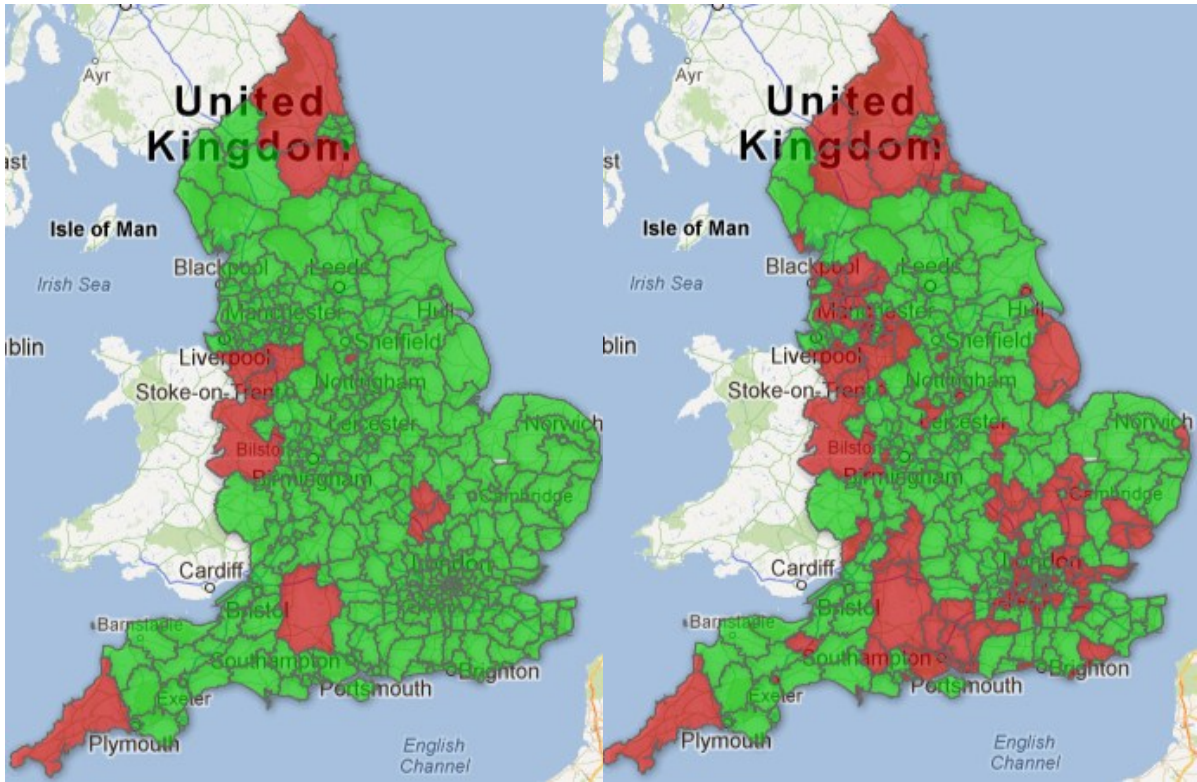


Figure 27: School types local authority areas

CY local authority areas 2005-2010

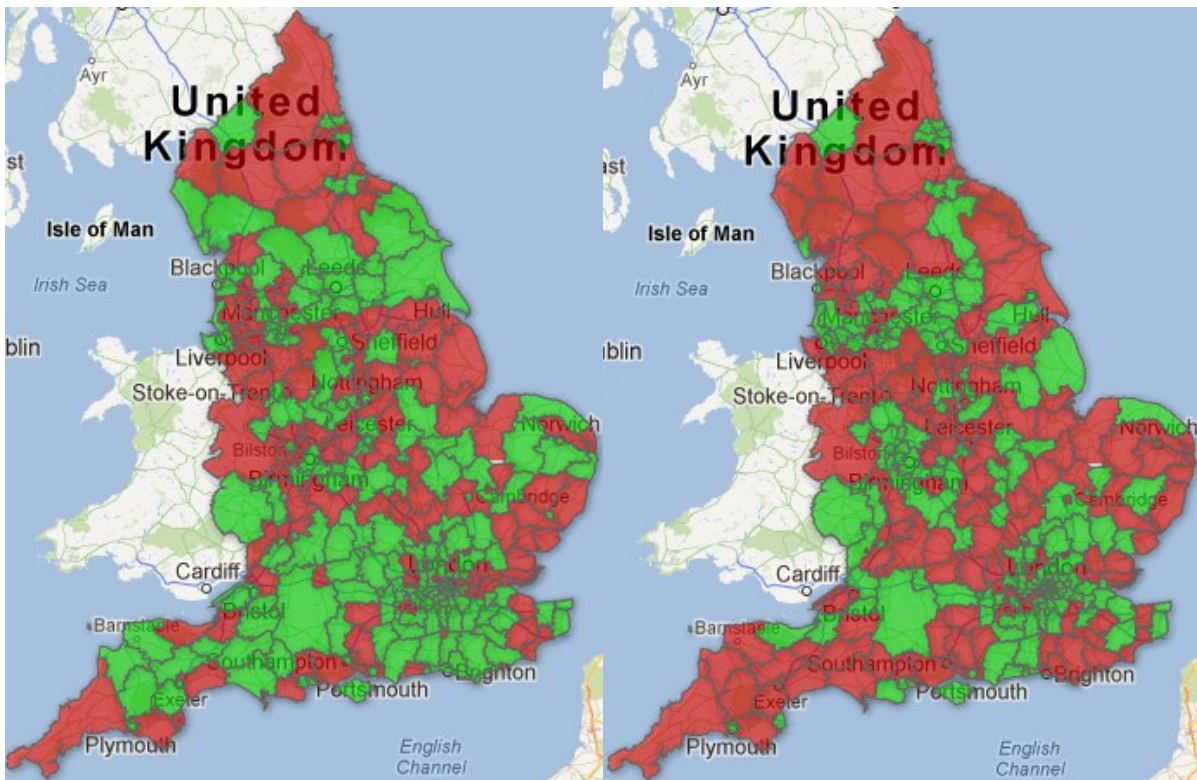


Figure 28: IND local authority areas

VA local authority areas 2005-2010

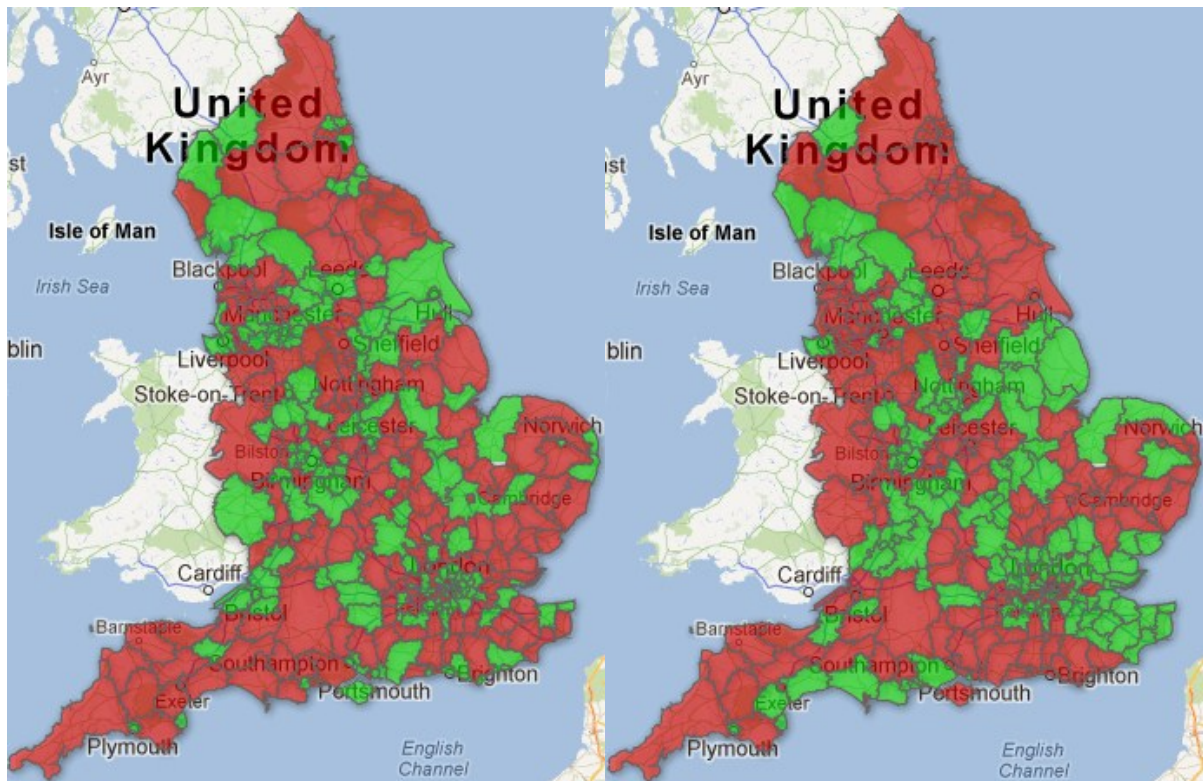


Figure 29: GFEC local authority areas

FD local authority areas 2005-2010

The correlation between the UKCI and other school types is not regarded separately as not enough records are concluded in the subgroups. As the figures show, the datasets for the different school types are incomparable, as the amount of school data records and regions they comprise, varies. To allow comparison between the school types, one dataset is created, integrating the Independent School dataset with the dataset of Community Schools, as those include the highest amount of school records and regions. All regions that are not comprised in both subgroups are deleted, to allow a comparable result concerning the different correlations within the two school type subgroups. The dataset comprises 200 regions, 2391 Community and 1623 Independent Schools.

Beside the sub datasets clustered by school types, datasets for each geographical entity are prepared. The segmentation of the overall dataset is shown in the pie chart below, denoting the number of regions and school records in each cluster.

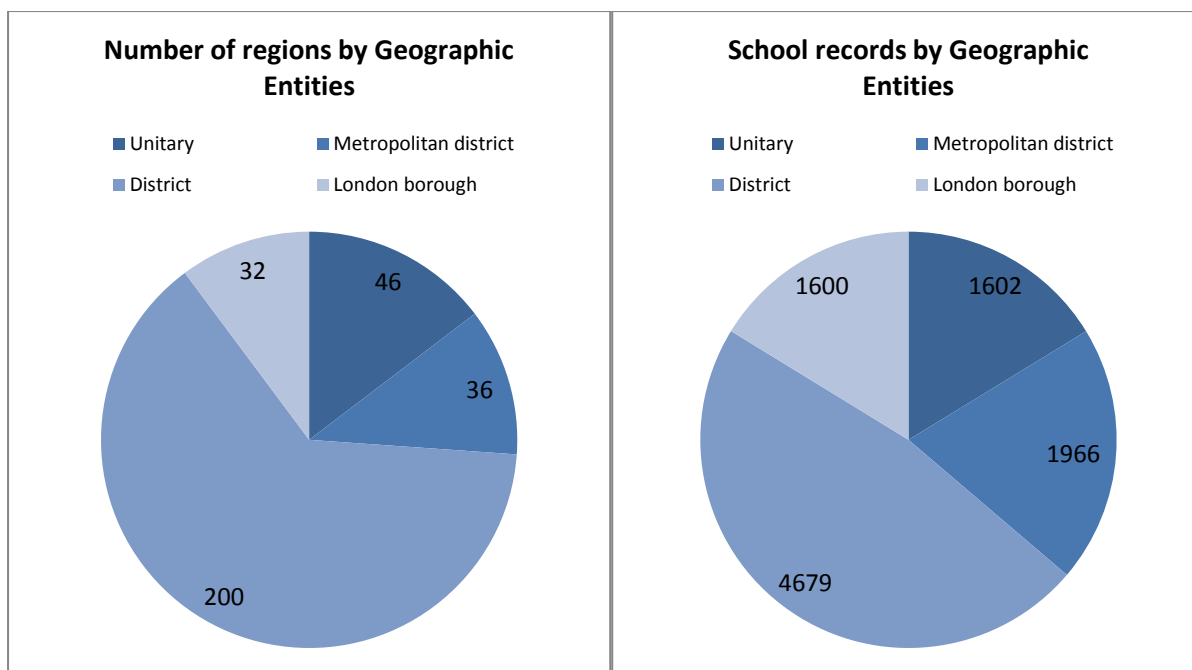


Figure 30: Segmentation by geographical entities

3.5. Data Analysis Model Deployment

Having cleaned and integrated the data, the data analysis model is deployed. The results are shown in the following chapters, structured by the research questions.

3.5.1. Correlation between A-Level performance and school properties

The correlation analysis between A-Level performance and school attributes is conducted pursuant to the breakdown analysis, depicted in Chapter 3.4.1. The calculated correlation statistics between the *A Level points per pupil* and the different school attributes are shown in the table below.

Attribute	A Level points per pupil
Number of pupils in 6 th form	Spearman: 0.09 Pearson: -0.15
School type	Ratio: 0.48 Square Ratio: 23%
Gender	Ratio: 0.36 Square Ratio: 13%
Admission policy	Ratio: 0.57 Square Ratio: 33%

Table 9: Correlation A Level points per pupil – school properties

Major findings resulting from these statistics are:

- No clear relation between the number of pupils in 6th form and school performance

- School performance is correlated to admission policy, school type and gender
- Especially the school's admission policy is a predictor for the school's performance

Next the dependencies between gender and school type and the performance of a school are investigated in more detail. Therefore, the correlation between gender and performance in the different subgroups clustered by school type and the correlation between school type and performance in the different subgroups clustered by gender are listed in the tables below.

School type	A Level points per pupil – gender correlation
Community School	Ratio: 0.15 Square Ratio: 2.4%
Independent School	Ratio: 0.29 Square Ratio: 8.6%
Foundation School	Ratio: 0.32 Square Ratio: 10%
Voluntary aided School	Ratio: 0.3 Square Ratio: 8.8%

Table 10: Correlation Gender – A-Level points per pupil in different school types

Gender	A Level points per pupil – school type correlation
Mixed	Ratio: 0.32 Square Ratio: 10%
Girls	Ratio: 0.29 Square Ratio: 8.4%
Boys	Ratio: 0.37 Square Ratio: 14%

Table 11: Correlation School type – A-Level points per pupil for different gender

Comparing the tables, Table 10 gives information about the correlation between gender and performance within the different school types, whereas Table 11 investigates on the correlation between school types and performance within the different gender schools. The findings give information about the predictability of a school's performance on the basis of its gender and school type. Major findings resulting from these statistics are:

- Gender does not have a significant influence on the performance of Community schools
- Foundation schools as well as Independent and Voluntary aided schools show correlation between their performance and gender policy
- Disregarding the school type, girls schools tend to score quite similar, whereas the performance of boys schools is correlated to the school type

Using the same methodology, the correlation between school type and admission policy and the performance of a school are investigated in more detail.

School type	A Level points per pupil – admission policy correlation
Community School	Ratio: 0.42 Square Ratio: 18%
Foundation School	Ratio: 0.68 Square Ratio: 46%
Voluntary aided School	Ratio: 0.57 Square Ratio: 33%

Table 12: Correlation Admission policy – A-Level points per pupil per school types

Admission policy	A Level points per pupil – school type correlation
Comprehensive	Ratio: 0.15 Square Ratio: 2.2%
Selective	Ratio: 0.17 Square Ratio: 3%
Modern	Ratio: 0.28 Square Ratio: 8%

Table 13: Correlation School type – A-Level points per pupil per admission policies

Table 12 gives information about the correlation between admission policy and performance within the different school types, whereas Table 13 investigates on the correlation between school type and performance within the applied admission policy. Major findings resulting from these statistics are:

- The performance of a distinct school type is correlated to the applied admission policy, especially for Foundation Schools
- The performance of the different school types, which apply a modern admission policy varies, whereas comprehensive and selective admission policies are a predictor for the performance of Community, Foundation and Voluntary aided schools
- The school type is strongly correlated to the admission policy, as the low correlation ratios indicates

The next tables show the correlation ratio, calculated between gender and admission policy with respect of the dependant variable *A-Level points per pupil*.

Gender	A Level points per pupil – admission policy correlation
Mixed	Ratio: 0.46 Square Ratio: 21%
Girls	Ratio: 0.72 Square Ratio: 52%
Boys	Ratio: 0.79 Square Ratio: 63%

Table 14: Correlation Admission policy – A-Level points per pupil per gender

Admission policy	A Level points per pupil – gender correlation
Comprehensive	Ratio: 0.06 Square Ratio: 0.3%
Selective	Ratio: 0.23 Square Ratio: 5%
Modern	Ratio: 0.09 Square Ratio: 0.9%

Table 15: Correlation Gender – A-Level points per pupil per admission policies

Major findings resulting from these statistics are:

- In general, the correlation between performance and admission policy is significantly higher for single-sex schools, for boys schools it is higher than for girls schools, due to huge differences between comprehensive and selective Voluntary aided and Foundation Schools
- Only for schools, applying a selective admission policy, a low correlation between its performance and gender policy can be assumed; compared to Table 10, this can be assigned to Foundation Schools, as those have the highest share of the data records for selective admission policy, as shown in Figure 19

Follow up the breakdown analysis, the correlation between *Pupils in 6th form* and *A-Level points per pupil* is analysed for the subgroups, clustered by school types, gender and school type and gender. The admission policy clusters are not further investigated in this context, as the subgroups would be too heterogeneous in respect of the *Pupils in 6th form* attribute. This table below shows the calculated coefficients for the different school types.

A-Level points per pupil - Pupils in 6th form	Spearman coefficient	Pearson coefficient
Community School	0.49	0.38
Independent School	0.49	0.41
Foundation School	0.47	0.41
Voluntary aided School	0.5	0.42
General Further Education College	0.27	0.31
Sixth Form College	0.31	0.31
Voluntary controlled School	0.45	0.36
Academy	0.35	0.36
Tertiary College	0.18	0.21

Table 16: Correlation Pupils in 6th form – A-Level points per pupil per school type

As the Spearman and Pearson coefficient show, there is a monotonic relation with linear characteristics between the A-Level performance of a school and its amount of pupils in 6th form. Especially in Community, Independent, Foundation and Voluntary aided and controlled schools the correlation between those two attributes is significant. To explain the correlation results, the distribution of the *Pupils in 6th form* value is visualized in the boxplot below.

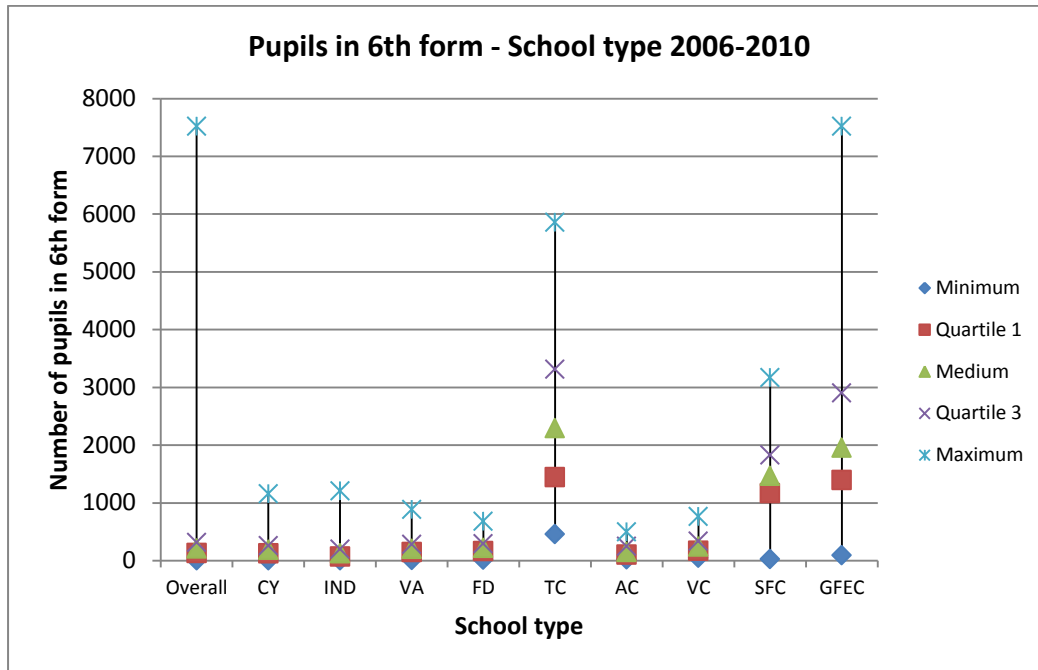


Figure 31: Boxplots Pupils in 6th form per school type 2006-2010

Comparing this data dispersion to the correlation coefficients, school types with high correlation coefficients show low dispersion with regard to their number of pupils in 6th form, whereas school type with high dispersion show low correlation.

Table 17 shows the coefficients for the subgroups clustered by gender.

A-Level points per pupil - Pupils in 6 th form	Spearman coefficient	Pearson coefficient
Mixed	0.19	-0.1
Girls	0.24	0.2
Boys	0.57	0.39

Table 17: Correlation Pupils in 6th form – A-Level points per pupil per gender

Especially boys schools show a strong monotonic relation with linear characteristics between their A-Level performance and number of pupils in 6th form. However, the correlation within mixed gender schools is not significant. This can also be explained by the value distribution of the attribute, as shown in the boxplots below.

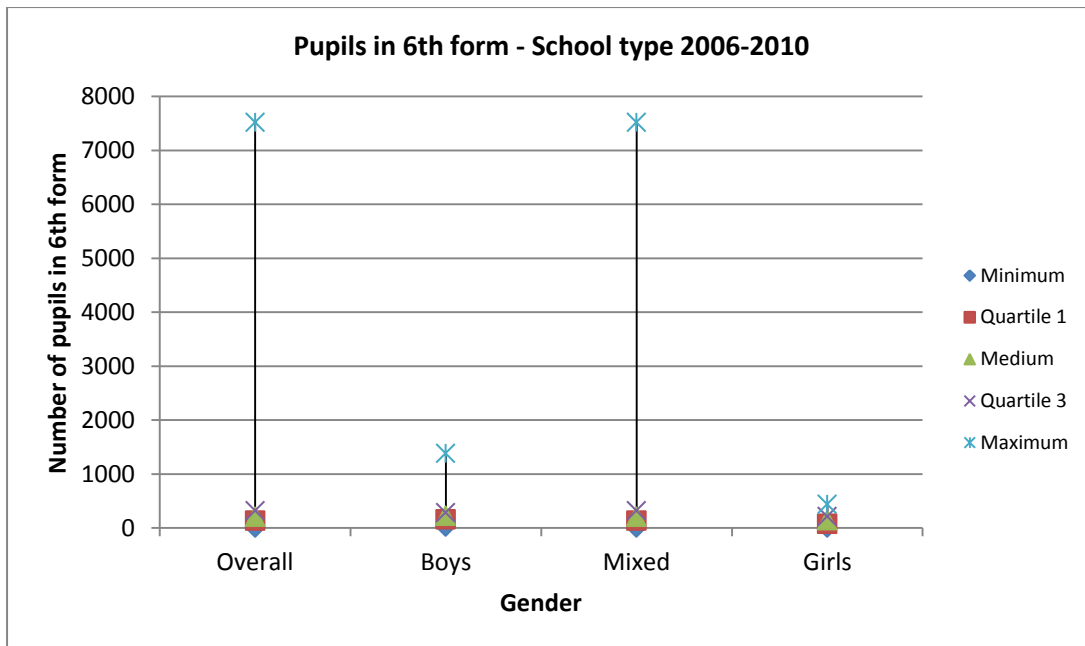


Figure 32: Boxplots Pupils in 6th form per gender 2006-2010

Next the correlation between A-Level performance and the number of pupils in 6th form is investigated, clustering the school types by gender.

A-Level points per pupil - Pupils in 6 th form	Mixed	Girls	Boys
Community School	Spearman: 0.49 Pearson: 0.39	Spearman: 0.38 Pearson: 0.36	Spearman: 0.74 Pearson: 0.7
Foundation School	Spearman: 0.44 Pearson: 0.4	Spearman: 0.47 Pearson: 0.56	Spearman: 0.74 Pearson: 0.76
Independent School	Spearman: 0.56 Pearson: 0.51	Spearman: 0.64 Pearson: 0.56	Spearman: 0.64 Pearson: 0.45
Voluntary aided School	Spearman: 0.53 Pearson: 0.47	Spearman: 0.34 Pearson: 0.25	Spearman: 0.21 Pearson: 0.27

Table 18: Correlation Pupils in 6th form – A-Level points per pupil school type - gender

Major findings resulting from these statistics are:

- In all cases, there is a relation between the attributes that tends to be linear
- The strongest correlation is revealed within boys schools, except for Voluntary aided boys schools
- Voluntary aided schools are the only school type where mixed schools have a higher correlation than single-sex schools
- The correlation is higher compared to the coefficients calculated for the school type cluster and gender cluster; the ratio of the coefficients of the different gender within this cluster is different than in the gender cluster

Comparing the calculated coefficients with the boxplots below, the relation to the dispersion of the attribute values cannot be approved, as for gender clusters and school type clusters.

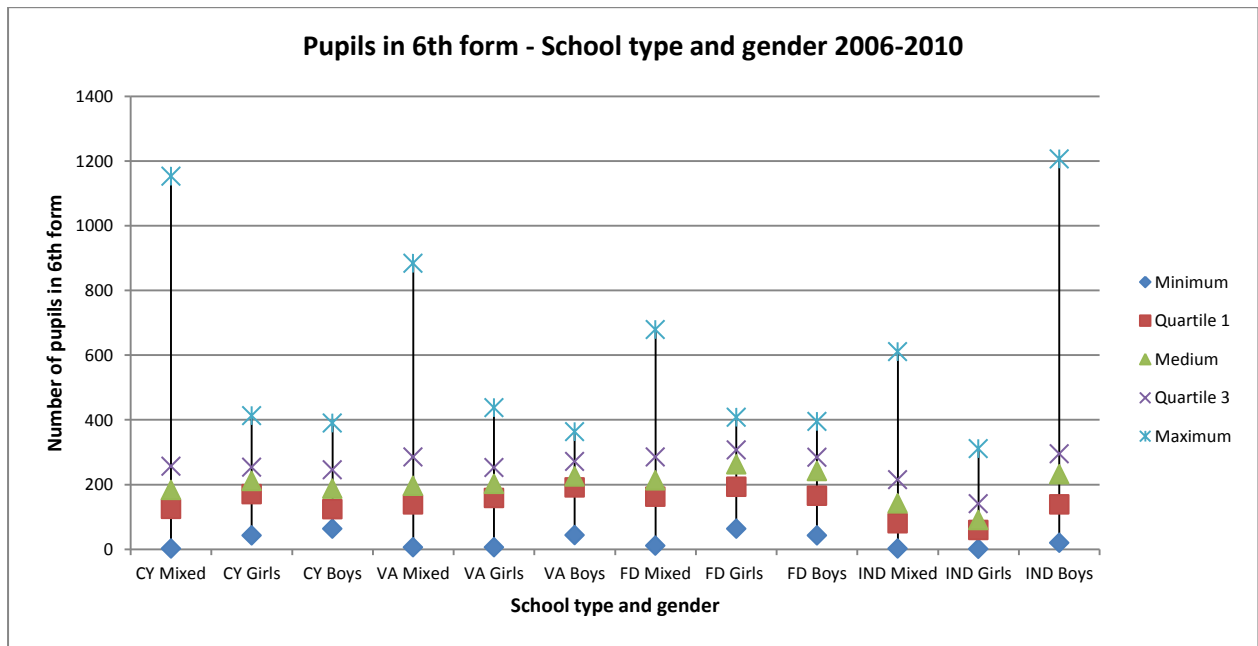


Figure 33: Boxplots Pupils in 6th form per gender and school type 2006-2010

Following, the explored linear correlations are visualized in a scatter plot matrix to evaluate and verify the significance of the calculated coefficients. The *A-Level points per pupil* are plotted on the horizontal axis and the *Pupils in 6th form* on the vertical axis. The maximum of the horizontal axis is 1500, the vertical axis has a maximum of 1200. Across the scattered data points a linear as well as a polynomial trend line is drawn, to highlight the distribution of the data and allow comparison of the different subgroups clustered by gender and school type.

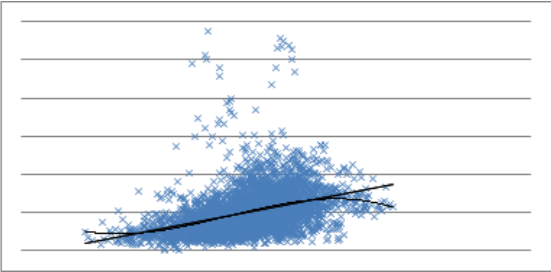
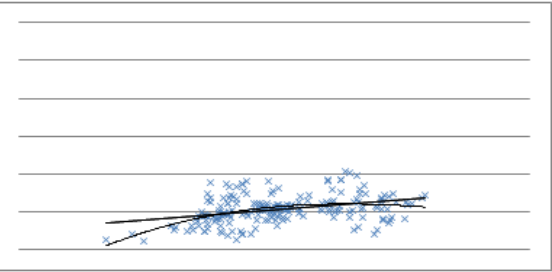
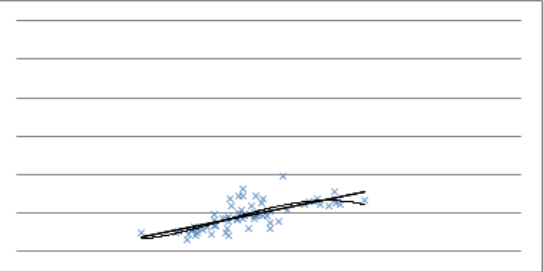
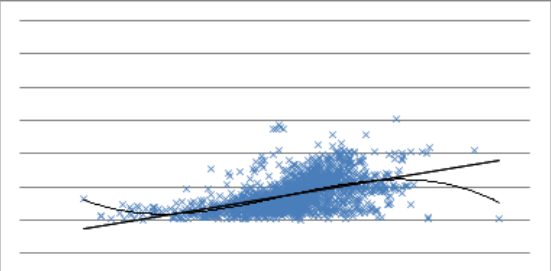
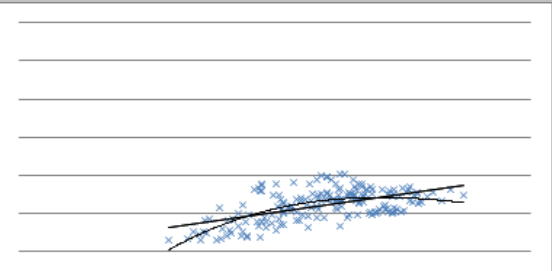
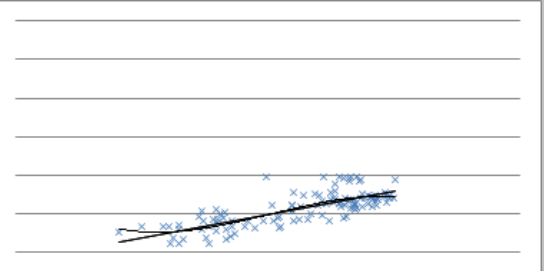
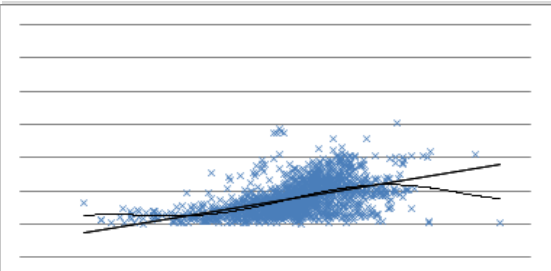
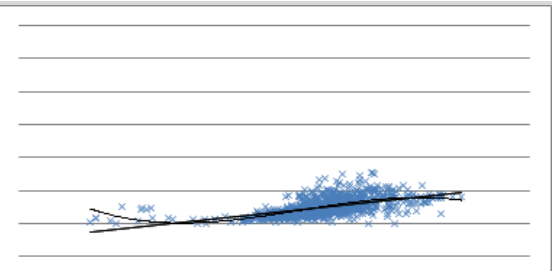
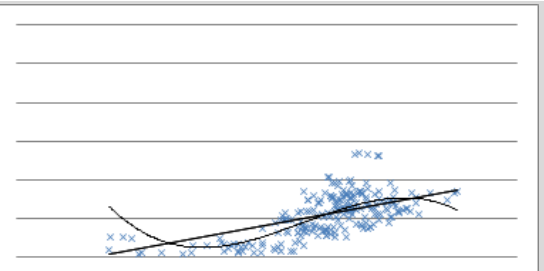
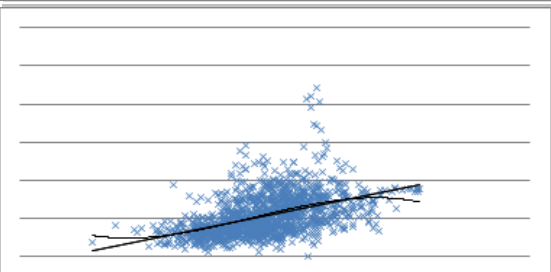
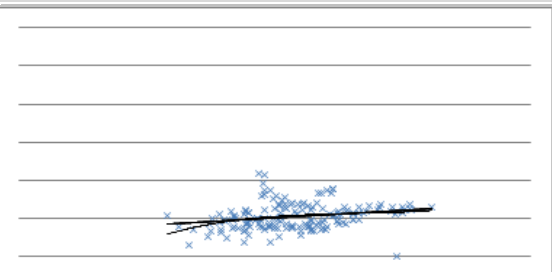
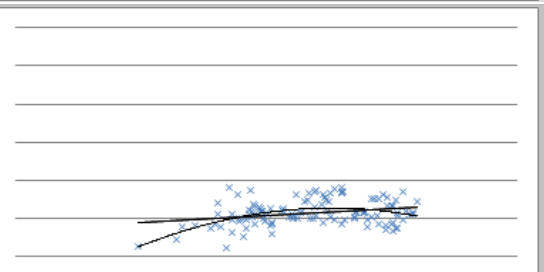
	Mixed	Girls	Boys
Community School			
Foundation School			
Independent School			
Voluntary aided School			

Table 19: Correlation Pupils in 6th form – A-Level points per pupil school type - gender

Major findings, resulting from an investigation of the scatter plots are:

- The significance of the coefficients is approved, as the calculated correlation between the *A-Level points per pupil* and *Pupils in 6th form* attribute within the different subgroups is visible and the linear relation is clearly presented by the linear trend line
- *A-Level points per pupil* increase with the *Pupils in 6th form* until a certain extent, but for schools with extremely high A-Level performance, the number of pupil in 6th forms tends to decrease, as the polynomial trend line shows

Having finalized the breakdown analysis, the investigations on the first research question are completed within the scope of this project. The next section analyses the data to find relevant information, concerning the second research question.

3.5.2. Correlation between A-Level performance and UK Competitive Index

First of all the correlation between performance and location of a school is calculated, using the correlation ratio. The ratio is calculated separately for each school year, to allow comparison to the subsequent correlation analysis between A-Level performance and UKCI.

Years	A Level points per pupil – school location correlation
2006	Ratio: 0.42; Square Ratio: 18%
2007	Ratio: 0.42; Square Ratio: 17%
2008	Ratio: 0.45 Square Ratio: 20%
2009	Ratio: 0.43 Square Ratio: 19%
2010	Ratio: 0.42 Square Ratio: 18%

Table 20: Correlation A-Level points per pupil – school location 2005-2010

The ratios show a potential correlation between the two attributes, however as the investigation is conducted on the data clustered by years, the correlation could also be reasoned by a correlation between the performance of a school and the school year. Therefore the associated correlation ratio is calculated:

Ratio: 0.09 Square Ratio: 1%

Although the year is not likely to influence the calculated correlation ratio between the performance of a school and its location, the correlation is not confirmed. The different local

authority area clusters are highly heterogeneous concerning for instance the amount of data records and the school types they comprise. As the school type is likely to be correlated with the performance of a school, as documented in the previous chapter, the correlation ratios in Table 20 are ambiguous. However an additional clustering by school types to get a more explicit result is not undertaken, as the clusters would comprise too few data records to get significant results for the ratio of variances.

Next, the correlation between A-Level performance of schools and the UK Competitive Index is investigated, deploying the project model as described in Chapter 3.3.1. The Pearson and Spearman's coefficients of the medians of the *A-Level points per pupil* attribute of each region and the corresponding Competitive Index, calculated for each pair of datasets of the different years, are shown in the matrix below. In contrast, the correlation between the averages over the years of the attributes of each region is shown in the left upper corner of the matrix. Similar matrices are created for Community, Independent, Foundation and Voluntary aided Schools as well as General Further Education Colleges as those school types comprise most data records and local authority areas as shown in Figure 25 and Figure 26.

Overall dataset $\bar{\rho}= 0.25$ $r= 0.18$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.25$ $r= 0.17$	$\rho= 0.24$ $r= 0.17$	$\rho= 0.21$ $r= 0.14$	$\rho= 0.23$ $r= 0.15$	$\rho= 0.23$ $r= 0.14$
A-Level 07	$\rho= 0.26$ $r= 0.19$	$\rho= 0.25$ $r= 0.19$	$\rho= 0.215$ $r= 0.15$	$\rho= 0.24$ $r= 0.17$	$\rho= 0.23$ $r= 0.16$
A-Level 08	$\rho= 0.27$ $r= 0.20$	$\rho= 0.27$ $r= 0.205$	$\rho= 0.24$ $r= 0.17$	$\rho= 0.27$ $r= 0.19$	$\rho= 0.25$ $r= 0.18$
A-Level 09	$\rho= 0.23$ $r= 0.165$	$\rho= 0.225$ $r= 0.17$	$\rho= 0.20$ $r= 0.14$	$\rho= 0.225$ $r= 0.15$	$\rho= 0.21$ $r= 0.14$
A-Level 10	$\rho= 0.265$ $r= 0.18$	$\rho= 0.265$ $r= 0.19$	$\rho= 0.23$ $r= 0.15$	$\rho= 0.27$ $r= 0.18$	$\rho= 0.25$ $r= 0.17$

Table 21: Correlation A-Level points per pupil - UK Competitive Index

CY Schools $\bar{\rho}= 0.18$ $r= 0.08$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.15$ $r= 0.08$	$\rho= 0.13$ $r= 0.07$	$\rho= 0.12$ $r= 0.06$	$\rho= 0.12$ $r= 0.04$	$\rho= 0.115$ $r= 0.04$
A-Level 07	$\rho= 0.14$ $r= 0.03$	$\rho= 0.155$ $r= 0.05$	$\rho= 0.12$ $r= 0.01$	$\rho= 0.12$ $r= -0.01$	$\rho= 0.105$ $r= -0.02$
A-Level 08	$\rho= 0.15$ $r= 0.06$	$\rho= 0.17$ $r= 0.08$	$\rho= 0.135$ $r= 0.05$	$\rho= 0.14$ $r= 0.04$	$\rho= 0.125$ $r= 0.03$
A-Level 09	$\rho= 0.17$ $r= 0.07$	$\rho= 0.185$ $r= 0.09$	$\rho= 0.145$ $r= 0.05$	$\rho= 0.14$ $r= 0.04$	$\rho= 0.135$ $r= 0.04$

A-Level 10	$\rho= 0.195$ $r= 0.12$	$\rho= 0.21$ $r= 0.14$	$\rho= 0.18$ $r= 0.11$	$\rho= 0.18$ $r= 0.10$	$\rho= 0.17$ $r= 0.10$
-------------------	----------------------------	---------------------------	---------------------------	---------------------------	---------------------------

Table 22: Correlation CY A-Level points per pupil - UK Competitive Index

IND Schools $\bar{\rho}= 0.09$ $r= 0.08$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.01$ $r= -0.02$	$\rho= 0.01$ $r= -0.01$	$\rho= 0.04$ $r= 0.02$	$\rho= 0.02$ $r= 0.01$	$\rho= 0.05$ $r= 0.02$
A-Level 07	$\rho= 0.05$ $r= 0.05$	$\rho= 0.05$ $r= 0.05$	$\rho= 0.07$ $r= 0.07$	$\rho= 0.07$ $r= 0.07$	$\rho= 0.09$ $r= 0.08$
A-Level 08	$\rho= 0.10$ $r= 0.07$	$\rho= 0.12$ $r= 0.09$	$\rho= 0.135$ $r= 0.11$	$\rho= 0.12$ $r= 0.11$	$\rho= 0.14$ $r= 0.12$
A-Level 09	$\rho= 0.005$ $r= 0.01$	$\rho= 0.02$ $r= 0.03$	$\rho= 0.04$ $r= 0.04$	$\rho= 0.04$ $r= 0.05$	$\rho= 0.05$ $r= 0.05$
A-Level 10	$\rho= 0.11$ $r= 0.08$	$\rho= 0.105$ $r= 0.09$	$\rho= 0.13$ $r= 0.11$	$\rho= 0.135$ $r= 0.14$	$\rho= 0.14$ $r= 0.14$

Table 23: Correlation IND A-Level points per pupil - UK Competitive Index

FD Schools $\bar{\rho}= 0.003$ $r=-0.03$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= -0.06$ $r= -0.07$	$\rho= -0.06$ $r= -0.08$	$\rho= -0.06$ $r= -0.07$	$\rho= -0.06$ $r= -0.07$	$\rho= -0.075$ $r= -0.11$
A-Level 07	$\rho= -0.04$ $r= -0.04$	$\rho= -0.06$ $r= -0.07$	$\rho= -0.06$ $r= -0.07$	$\rho= -0.04$ $r= -0.06$	$\rho= -0.06$ $r= -0.09$
A-Level 08	$\rho= 0.05$ $r= 0.02$	$\rho= 0.025$ $r= -0.01$	$\rho= 0.03$ $r= -0.01$	$\rho= 0.05$ $r= 0.01$	$\rho= 0.03$ $r= -0.02$
A-Level 09	$\rho= 0.05$ $r= 0.03$	$\rho= 0.04$ $r= 0.001$	$\rho= 0.03$ $r= -0.01$	$\rho= 0.06$ $r= 0.01$	$\rho= 0.03$ $r= -0.02$
A-Level 10	$\rho= 0.09$ $r= 0.02$	$\rho= 0.07$ $r= -0.01$	$\rho= 0.05$ $r= -0.02$	$\rho= 0.07$ $r= 0.001$	$\rho= 0.06$ $r= -0.02$

Table 24: Correlation FD A-Level points per pupil - UK Competitive Index

VA Schools $\bar{\rho}= 0.02$ $r= -0.01$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.02$ $r= -0.02$	$\rho= 0.01$ $r= -0.03$	$\rho= -0.02$ $r= -0.06$	$\rho= -0.03$ $r= -0.07$	$\rho= -0.035$ $r= -0.07$
A-Level 07	$\rho= 0.05$ $r= 0.01$	$\rho= 0.05$ $r= 0.01$	$\rho= 0.01$ $r= -0.015$	$\rho= 0.02$ $r= -0.02$	$\rho= 0.01$ $r= -0.02$
A-Level 08	$\rho= 0.12$ $r= 0.08$	$\rho= 0.12$ $r= 0.075$	$\rho= 0.08$ $r= 0.05$	$\rho= 0.06$ $r= 0.03$	$\rho= 0.05$ $r= 0.03$
A-Level 09	$\rho= 0.06$ $r= 0.03$	$\rho= 0.05$ $r= 0.02$	$\rho= 0.01$ $r= -0.01$	$\rho= 0.01$ $r= -0.02$	$\rho= -0.005$ $r= -0.03$
A-Level 10	$\rho= 0.065$ $r= 0.02$	$\rho= 0.05$ $r= 0.01$	$\rho= 0.02$ $r= -0.015$	$\rho= 0.02$ $r= -0.02$	$\rho= 0.02$ $r= -0.02$

Table 25: Correlation VA A-Level points per pupil - UK Competitive Index

GFEC $\bar{\rho} = 0.13$ $r = 0.07$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho = 0.19$ $r = 0.08$	$\rho = 0.175$ $r = 0.08$	$\rho = 0.15$ $r = 0.05$	$\rho = 0.19$ $r = 0.08$	$\rho = 0.18$ $r = 0.07$
A-Level 07	$\rho = 0.13$ $r = 0.09$	$\rho = 0.115$ $r = 0.08$	$\rho = 0.09$ $r = 0.06$	$\rho = 0.11$ $r = 0.07$	$\rho = 0.10$ $r = 0.06$
A-Level 08	$\rho = 0.16$ $r = 0.08$	$\rho = 0.16$ $r = 0.09$	$\rho = 0.13$ $r = 0.06$	$\rho = 0.17$ $r = 0.08$	$\rho = 0.16$ $r = 0.07$
A-Level 09	$\rho = 0.1$ $r = 0.05$	$\rho = 0.10$ $r = 0.06$	$\rho = 0.07$ $r = 0.03$	$\rho = 0.11$ $r = 0.07$	$\rho = 0.11$ $r = 0.06$
A-Level 10	$\rho = 0.08$ $r = 0.04$	$\rho = 0.08$ $r = 0.05$	$\rho = 0.05$ $r = 0.02$	$\rho = 0.10$ $r = 0.05$	$\rho = 0.1$ $r = 0.04$

Table 26: Correlation GFEC A-Level points per pupil - UK Competitive Index

Referring to the coefficients, a correlation between the attributes within the whole dataset might be assumed, as well as in the subgroups of Community Schools, Independent Schools and General Further Education Colleges. For those datasets the distribution of the data might have linear tendencies. Comparing the different year combinations of the paired data, no patterns regarding the time gap of interaction can be suggested.

To get a clearer understanding of the results, the correlation in the whole dataset as well as in the subgroups clustered by school types is visualized in the scatterplots below. The average UK Competitive Index of a region over the years 2006 to 2010 - excluding 2007 - is plotted on the horizontal axis, the average *A-Level points per pupil* of a region over the years 2006 to 2010 - excluding 2007 - on the vertical axis. Subsequently, the scatter charts visualizes the correlation coefficients, shown in the left upper corner of the previous tables. Across the scattered data points a linear trend line is drawn, to highlight the distribution of the data and allow comparison of the correlation within the different school type subgroups.

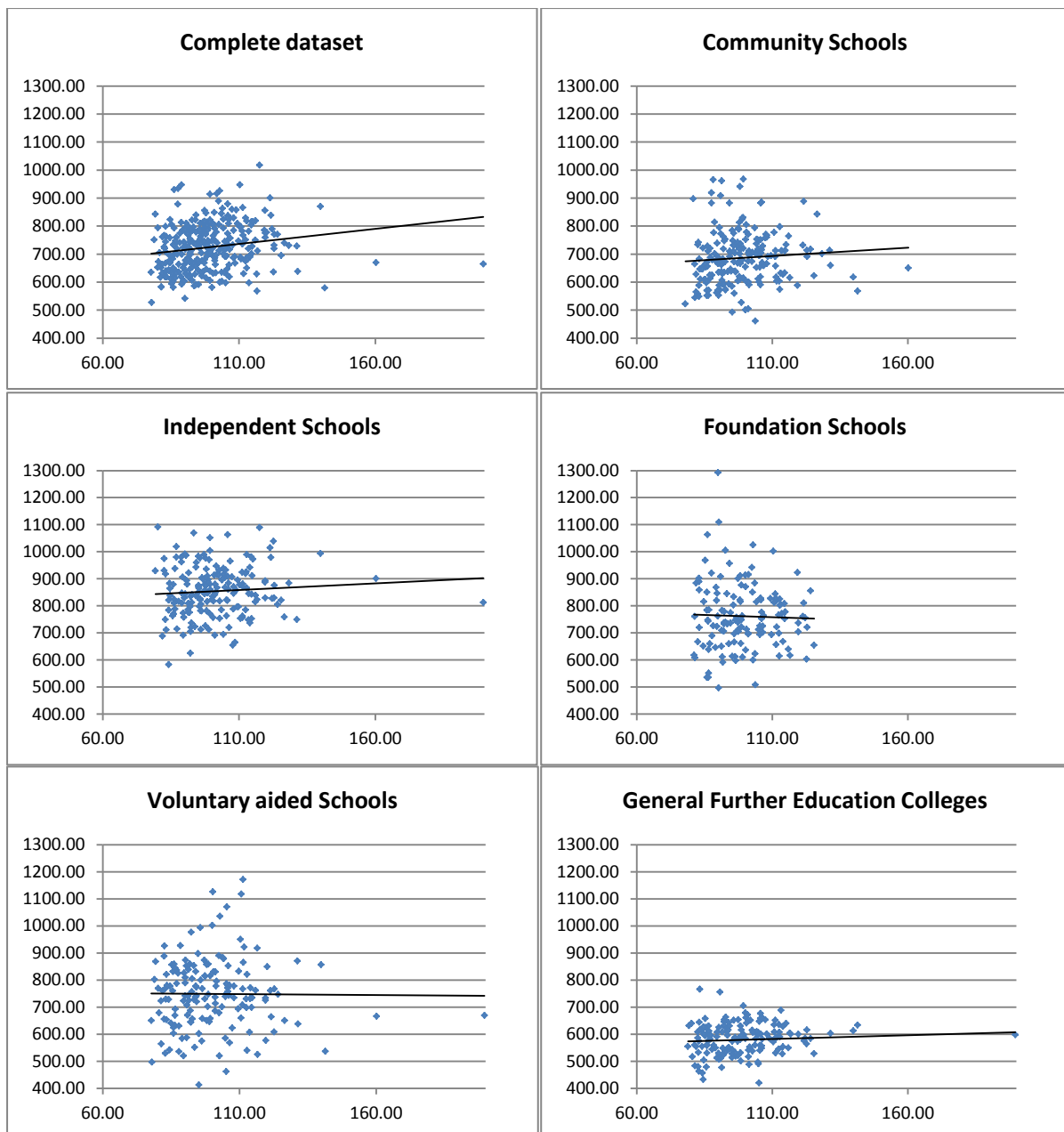


Figure 34: School type clusters correlation A-Level points per pupil – UKCI

These scatter plots support the conclusions drawn from the results of the coefficient calculations:

- The data for Foundation and Voluntary aided Schools are scattered and show no linearity
- The data cloud in the scatter chart of Independent and Community Schools is similar to the one of the overall dataset, but shows now clear correlation
- The data of General Further Education Schools is concentrated, but also no clear correlation is shown

To investigate deeper into the correlation between UKCI and A-Level performance in the subgroups clustered by school type, the coefficients for the integrated dataset comprising Independent and Community Schools - described in Chapter 3.4.2 - are calculated and shown in the tables below.

CY Schools $\bar{\rho}= 0.22$ $r= 0.13$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.17$ $r= 0.11$	$\rho= 0.15$ $r= 0.09$	$\rho= 0.13$ $r= 0.08$	$\rho= 0.13$ $r= 0.08$	$\rho= 0.13$ $r= 0.07$
A-Level 07	$\rho= 0.20$ $r= 0.11$	$\rho= 0.21$ $r= 0.13$	$\rho= 0.18$ $r= 0.10$	$\rho= 0.16$ $r= 0.07$	$\rho= 0.15$ $r= 0.06$
A-Level 08	$\rho= 0.22$ $r= 0.10$	$\rho= 0.25$ $r= 0.14$	$\rho= 0.20$ $r= 0.11$	$\rho= 0.18$ $r= 0.07$	$\rho= 0.18$ $r= 0.06$
A-Level 09	$\rho= 0.24$ $r= 0.13$	$\rho= 0.26$ $r= 0.17$	$\rho= 0.21$ $r= 0.13$	$\rho= 0.20$ $r= 0.12$	$\rho= 0.20$ $r= 0.11$
A-Level 10	$\rho= 0.195$ $r= 0.23$	$\rho= 0.25$ $r= 0.17$	$\rho= 0.21$ $r= 0.14$	$\rho= 0.20$ $r= 0.13$	$\rho= 0.20$ $r= 0.12$

Table 27: Correlation CY A-Level points per pupil - UK Competitive Index

IND Schools $\bar{\rho}= 0.07$ $r= 0.07$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.02$ $r= -0.01$	$\rho= 0.01$ $r= -0.01$	$\rho= 0.05$ $r= 0.02$	$\rho= 0.02$ $r= 0.01$	$\rho= 0.04$ $r= 0.02$
A-Level 07	$\rho= 0.03$ $r= 0.02$	$\rho= 0.02$ $r= 0.01$	$\rho= 0.04$ $r= 0.04$	$\rho= 0.02$ $r= 0.04$	$\rho= 0.05$ $r= 0.05$
A-Level 08	$\rho= 0.05$ $r= 0.03$	$\rho= 0.07$ $r= 0.06$	$\rho= 0.10$ $r= 0.08$	$\rho= 0.07$ $r= 0.08$	$\rho= 0.08$ $r= 0.08$
A-Level 09	$\rho= -0.02$ $r= -0.001$	$\rho= -0.003$ $r= 0.02$	$\rho= 0.02$ $r= 0.04$	$\rho= 0.004$ $r= 0.04$	$\rho= 0.01$ $r= 0.04$
A-Level 10	$\rho= 0.09$ $r= 0.11$	$\rho= 0.11$ $r= 0.13$	$\rho= 0.135$ $r= 0.16$	$\rho= 0.135$ $r= 0.15$	$\rho= 0.11$ $r= 0.14$

Table 28: Correlation IND A-Level points per pupil - UK Competitive Index

The coefficients are quite similar to the results of the original data set, shown in Table 22 and Table 23. Concluding the results, Community Schools show higher correlation between their A-Level performance and the local UKCI than Independent Schools, regarding the 122 local authority areas, applied in the dataset.

Next the tables are established for each subgroup, clustered by geographic entity, to investigate on the correlation between A-Level performance and UKCI within different local governments. In England four entities are differentiated (Google, 2012):

London boroughs, metropolitan districts, rural districts and unitary authorities

London $\bar{\rho}= 0.03$ $r= 0.09$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.08$ $r= 0.08$	$\rho= 0.09$ $r= 0.08$	$\rho= -0.01$ $r= 0.05$	$\rho= 0.03$ $r= 0.04$	$\rho= 0.01$ $r= 0.02$
A-Level 07	$\rho= 0.09$ $r= 0.05$	$\rho= 0.10$ $r= 0.05$	$\rho= 0.002$ $r= 0.01$	$\rho= 0.09$ $r= 0.04$	$\rho= 0.07$ $r= 0.02$
A-Level 08	$\rho= 0.15$ $r= 0.02$	$\rho= 0.17$ $r= 0.01$	$\rho= 0.09$ $r= -0.02$	$\rho= 0.21$ $r= 0.06$	$\rho= 0.18$ $r= 0.03$
A-Level 09	$\rho= 0.125$ $r= 0.02$	$\rho= 0.09$ $r= 0.002$	$\rho= 0.02$ $r= -0.015$	$\rho= 0.19$ $r= 0.09$	$\rho= 0.16$ $r= 0.06$
A-Level 10	$\rho= 0.09$ $r= -0.02$	$\rho= 0.07$ $r= -0.03$	$\rho= -0.01$ $r= -0.06$	$\rho= 0.11$ $r= 0.09$	$\rho= 0.085$ $r= 0.07$

Table 29: Correlation London boroughs A-Level points per pupil - UKCI

Metropolitan $\bar{\rho}= 0.34$ $r= 0.48$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.29$ $r= 0.44$	$\rho= 0.26$ $r= 0.37$	$\rho= 0.28$ $r= 0.41$	$\rho= 0.36$ $r= 0.40$	$\rho= 0.34$ $r= 0.42$
A-Level 07	$\rho= 0.40$ $r= 0.51$	$\rho= 0.38$ $r= 0.46$	$\rho= 0.38$ $r= 0.48$	$\rho= 0.43$ $r= 0.47$	$\rho= 0.39$ $r= 0.47$
A-Level 08	$\rho= 0.35$ $r= 0.51$	$\rho= 0.32$ $r= 0.44$	$\rho= 0.32$ $r= 0.48$	$\rho= 0.39$ $r= 0.46$	$\rho= 0.38$ $r= 0.49$
A-Level 09	$\rho= 0.43$ $r= 0.54$	$\rho= 0.42$ $r= 0.47$	$\rho= 0.50$ $r= 0.04$	$\rho= 0.45$ $r= 0.47$	$\rho= 0.42$ $r= 0.48$
A-Level 10	$\rho= 0.36$ $r= 0.51$	$\rho= 0.35$ $r= 0.45$	$\rho= 0.32$ $r= 0.48$	$\rho= 0.45$ $r= 0.50$	$\rho= 0.44$ $r= 0.52$

Table 30: Correlation Metropolitan districts A-Level points per pupil - UKCI

Unitary $\bar{\rho}= 0.095$ $r= 0.12$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho= 0.185$ $r= 0.17$	$\rho= 0.20$ $r= 0.17$	$\rho= 0.18$ $r= 0.17$	$\rho= 0.19$ $r= 0.17$	$\rho= 0.20$ $r= 0.18$
A-Level 07	$\rho= 0.22$ $r= 0.16$	$\rho= 0.23$ $r= 0.15$	$\rho= 0.21$ $r= 0.16$	$\rho= 0.23$ $r= 0.16$	$\rho= 0.23$ $r= 0.17$
A-Level 08	$\rho= 0.20$ $r= 0.20$	$\rho= 0.22$ $r= 0.20$	$\rho= 0.20$ $r= 0.19$	$\rho= 0.18$ $r= 0.17$	$\rho= 0.18$ $r= 0.18$
A-Level 09	$\rho= 0.0004$ $r= 0.04$	$\rho= 0.02$ $r= 0.04$	$\rho= 0.01$ $r= 0.04$	$\rho= -0.004$ $r= 0.03$	$\rho= -0.04$ $r= 0.03$
A-Level 10	$\rho= 0.14$ $r= 0.07$	$\rho= 0.15$ $r= 0.08$	$\rho= 0.135$ $r= 0.08$	$\rho= 0.13$ $r= 0.06$	$\rho= 0.13$ $r= 0.07$

Table 31: Correlation Unitary A-Level points per pupil – UKCI

District $\bar{\rho} = 0.33$ $r = 0.33$	UKCI 05	UKCI 06	UKCI 08	UKCI 09	UKCI 10
A-Level 06	$\rho = 0.31$ $r = 0.30$	$\rho = 0.33$ $r = 0.31$	$\rho = 0.29$ $r = 0.32$	$\rho = 0.33$ $r = 0.32$	$\rho = 0.33$ $r = 0.32$
A-Level 07	$\rho = 0.29$ $r = 0.29$	$\rho = 0.28$ $r = 0.29$	$\rho = 0.25$ $r = 0.30$	$\rho = 0.29$ $r = 0.30$	$\rho = 0.28$ $r = 0.29$
A-Level 08	$\rho = 0.32$ $r = 0.32$	$\rho = 0.34$ $r = 0.34$	$\rho = 0.30$ $r = 0.34$	$\rho = 0.33$ $r = 0.34$	$\rho = 0.31$ $r = 0.32$
A-Level 09	$\rho = 0.29$ $r = 0.29$	$\rho = 0.30$ $r = 0.31$	$\rho = 0.27$ $r = 0.29$	$\rho = 0.30$ $r = 0.29$	$\rho = 0.29$ $r = 0.28$
A-Level 10	$\rho = 0.31$ $r = 0.30$	$\rho = 0.33$ $r = 0.32$	$\rho = 0.29$ $r = 0.30$	$\rho = 0.32$ $r = 0.30$	$\rho = 0.30$ $r = 0.29$

Table 32: Correlation District A-Level points per pupil - UKCI

Referring to the coefficients, a significant correlation between the attributes within Metropolitan district and rural districts might be assumed. For those datasets, the distribution of the data might have linear tendencies. Comparing the different year combinations of the paired data, no patterns regarding the time gap of interaction can be suggested.

To get a clearer understanding of the results, the correlation in the subgroups, clustered by geographical entity is visualized in the scatterplots below. The average UK Competitive Index of a region over the years 2006 to 2010 - excluding 2007 - is plotted on the horizontal axis and the average *A-Level points per pupil* of a region over the years 2006 to 2010 - excluding 2007 - on the vertical axis. Across the scattered data points, a linear trend line is drawn to highlight the distribution of the data and allow comparison of the different clusters.

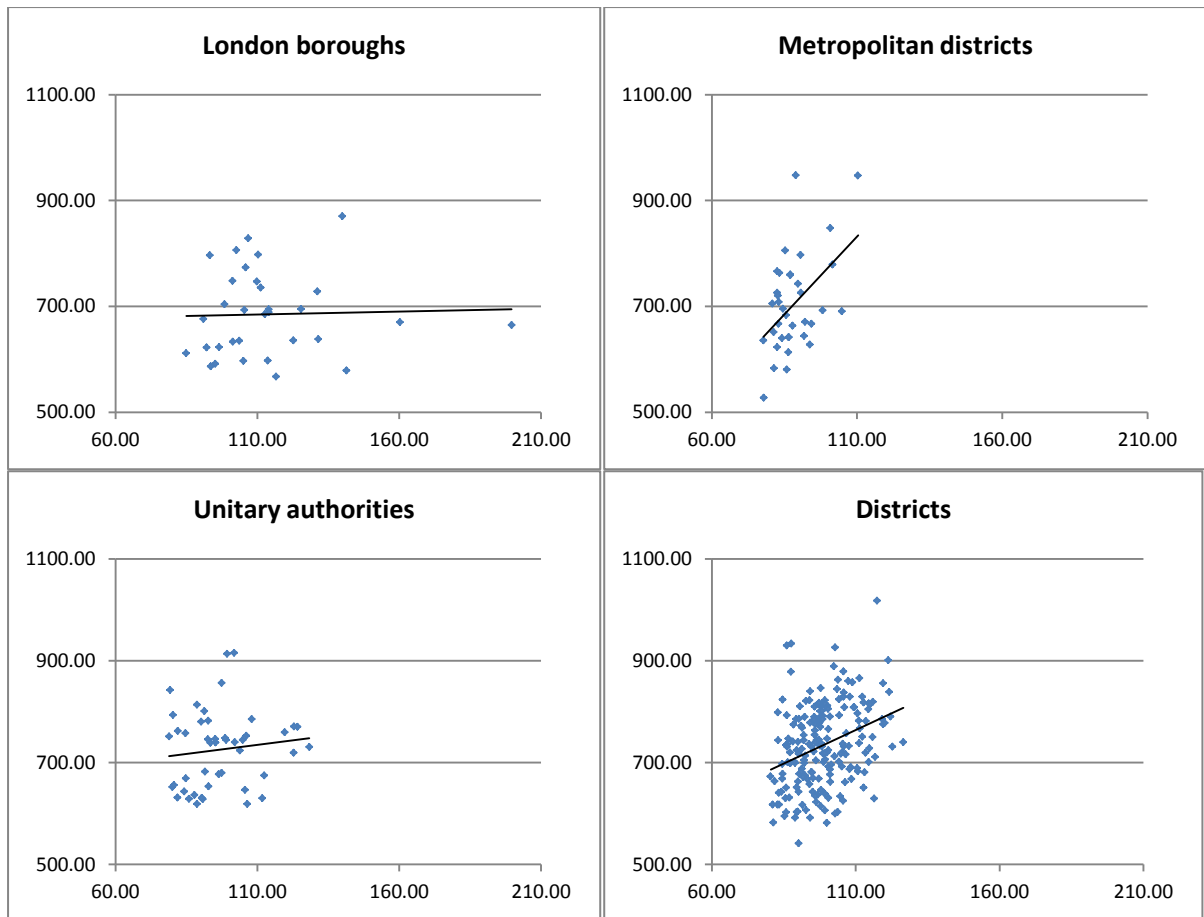


Figure 35: Geographic entity clusters Correlation A-Level points per pupil – UKCI

These scatter plots support the conclusions, drawn from the results of the coefficient calculations:

- The data values for London boroughs and unitary authorities are scattered and show no linearity
- Metropolitan and rural districts show monotonic correlation between local average A-Level performance and UKCI

After analysing the datasets, clustered by school type and geographic entity, the investigation on the correlation between the socio-economic situation of a region and its schools' average performance in A-Levels is finalized in the context of this dissertation project. Before the conducted data analysis is concluded by evaluating its results, the major findings are summarized.

3.5.3. Summarization of results

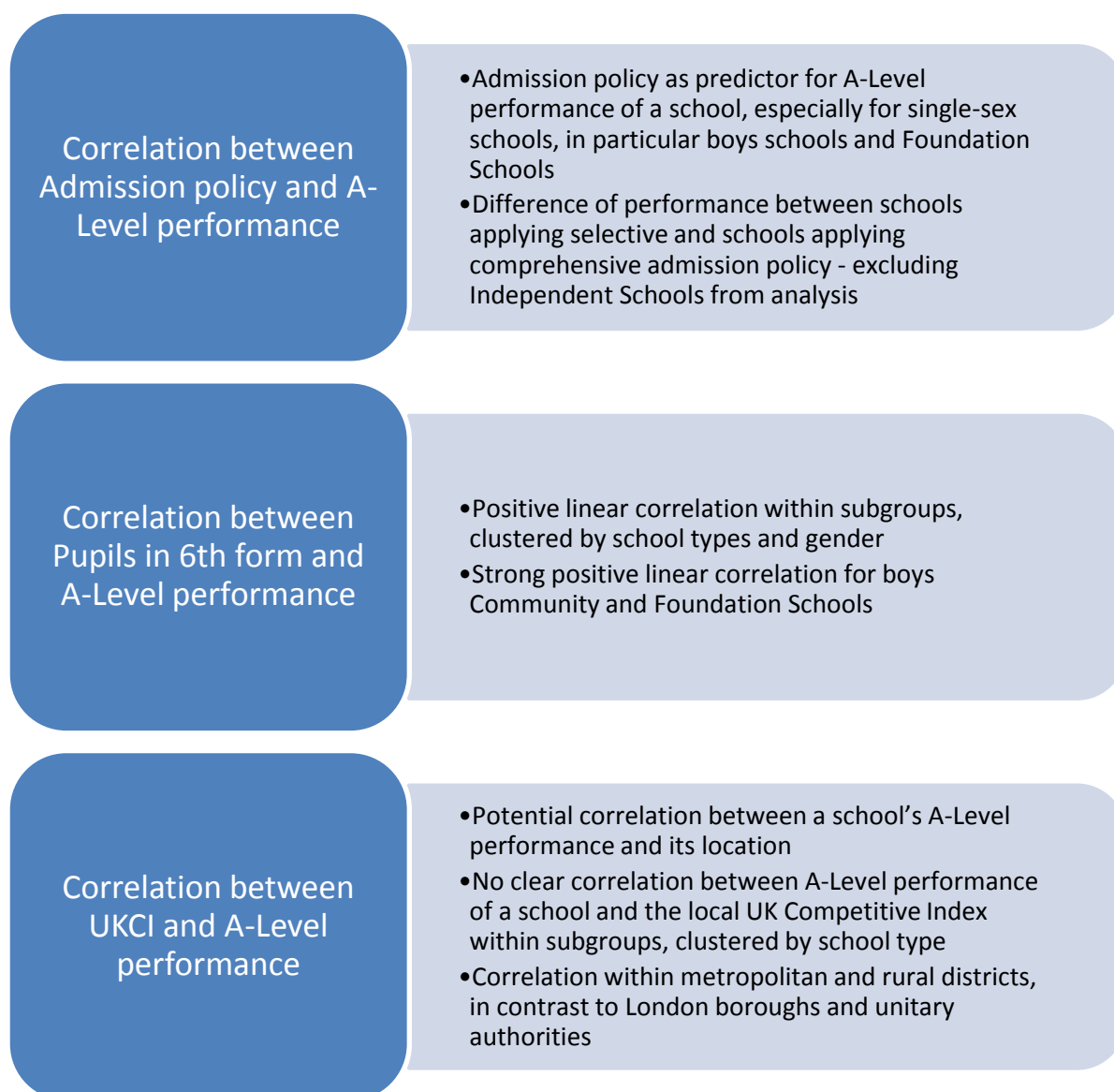


Figure 36: Summary of major findings

4. Conclusion

Concluding this data analysis project, the mission and objectives, as stated in the introduction chapters of this dissertation, are accomplished. Within the limited time frame, set out for this research, an insight in the large datasets of school and student data is gained, using an exploratory visual approach to data analysis. During this process, interesting patterns with respect to school performance are discovered, resulting in two detailed research questions that require an in depth investigation. Additional data records are integrated, to analyse correlations between socio-economic factors and school performance on local level and an

appropriate data model is generated. The results of the data analysis disclose interesting and in some cases significant correlations between school properties and school performance. However, the significance of the results of the correlation analysis between A-Level performance and the UK Competitive index is limited, due to the quality of the analysed data. As the UKCI is composed of six different socio-economic statistics, that are weighted based on a three factor model, it is likely to be inaccurate. Furthermore it is only available for a period of six years, including one gap year. Therefore, a profound time series analysis between the UKCI and A-Level performance of a region, implying auto- and cross correlation, cannot be conducted. The model developed in the context of this research project is not capable to comprise the entire complexity of correlations and disturbing external factors over the years.

However, the disclosed correlations between the A-Level performance of a school and its properties as well as the revealed potential relation to the local socio-economic situation, provides the grounds for further research. The developed data analysis model can easily be applied to further numeric and nominal data attributes of the school and pupil data sets, to discover further correlations. The most obvious option would be the expansion of the correlation analysis on the GCSE and KS2 performance tables. Besides the separated analysis of their performance tables, the inter-correlation between KS2, GCSE and A-Level performance could be examined, conducting a profound time series analysis that takes into account the findings of this research.

A more complex extension of the conducted research could concentrate on the development of an accurate indicator for the socio-economic situation of local authority areas. The Office for National Statistics (Office for National Statistics, 2012) provides various datasets extending back to 2005, comprising socio-economic statistics – for instance unemployment rates, immigration rates, crime rates or income rates - broken down to the level of local authority areas. After a correlation analysis of those factors, a model similar to the UKCI could be established and applied on the datasets. Having a representative index over an appropriate period of time, a profound time series analysis could be conducted to reveal correlations between the socio-economic wellbeing of an area and its educational achievements.

Bibliography

- Bergdahl, M., Ehling, M., Elvers, E., Földesi, E., Körner, T., Kron, A., et al. (2007). Handbook on Data Quality Assessment Methods and Tools. *Handbook on Data Quality Assessment Methods and Tools* (pp. 9-10). Wiesbaden: European Commission.
- Berthold, M., Borgelt, C., Hoepfner, F., & Klawoon, F. (2010). *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*. London: Springer.
- Department for Education. (2012). *Technical annex*. Retrieved July 05, 2012, from http://www.education.gov.uk/performance/tables/pilot16_05/annex.shtml
- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, 43(1), 50-54 .
- Google. (2012). *Welcome to Fusion Tables*. Retrieved June 09, 2012, from <http://support.google.com/fusiontables/bin/answer.py?hl=en&answer=2571232>
- Gupta, S. (2009). Business statistics. Jaipur: Sultan Chand & Sons.
- Huggins, R. (2003). Creating a UK Competitive Index: Regional and Local Benchmarking. *Regional Studies*, 37(1), 89-96.
- Mirkin, B. (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*. New York: Springer.
- Office for National Statistics. (2012). *Datasets and reference tables*. Retrieved June 02, 2012, from <http://www.ons.gov.uk/ons/datasets-and-tables/index.html>
- Park, E., & Lee, Y. (2001). Estimates of Standard Deviation of Spearman's Rank Correlation Coefficients with Dependent Observations. *Communications In Statistics: Simulation & Computation*, 30(1), 129–142.
- QlikTech International AB. (2011). *QlikView Reference Manual*. Lund: QlikTech International AB.
- Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), 59-66.
- Seale, C. (2004). *Researching Society and Culture*. London: SAGE Publications Ltd.
- The Good Schools Guide. (2012). *The Good Schools Guide*. Retrieved May 25, 2012, from <http://www.goodschoolsguide.co.uk/>
- Tukey, J. (1977). *Exploratory Data Analysis*. Mass: Addison-Wesley Pub. Co.
- Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A Framework of Analysis of Data Quality Research. *IEEE Transactions on knowledge and data engineering*, 7(4), 623 - 640.

Appendices

Quartiles A-Level points per pupil 2006 to 2009

A-Level points per pupil	Red	Pink	Yellow	Green
2006	<=585	585<=701	701<=813	813<
2007	<=594	594<=703	703<=824	824<
2008	<=616	616<=722	722<=831	831<
2009	<=621	621<=725	725<=832	832<

Table 33: Quartiles A-Level points per pupil 2006 to 2009

School data in regional context 2006 to 2009

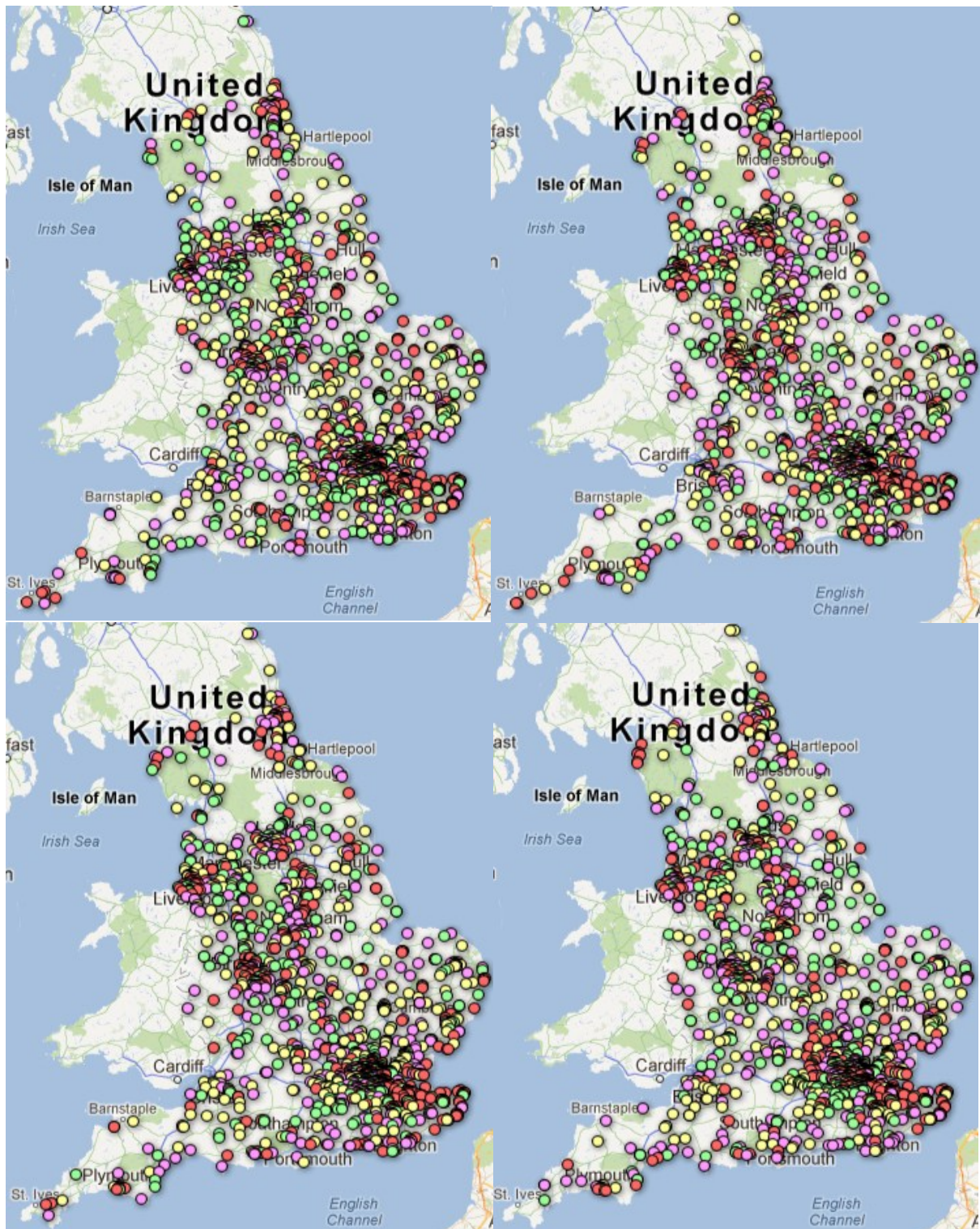


Figure 37: School data in regional context 2006 to 2009