CAMBRIDGE ASSESSMENT

Using association rules to understand subject choice at AS/A level

Tom Sutch

Cambridge Assessment Research Report

4 December 2015

University of Cambridge
Local Examinations Syndicate

Author contact details:

ARD Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG

sutch.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk/

How to cite this publication:

Sutch, T. (2015). *Using association rules to understand subject choice at AS/A level*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Contents

Summary iv
1 Introduction 1
1.1 Context 1
1.2 Association rules 1
1.2.1 Explanation 1
1.2.2 Application in educational research 3
1.2.3 Other data mining techniques on similar data 4
2 Aims of this research 4
3 Data/methods 4
4 Results 6
4.1 Results: basic rules 6
4.2 Results: demographic information included in transactions 12
4.3 Results: contrasting rules 18
5 Discussion 23
References 25

Summary

Association rule analysis is a data mining technique, with origins in retail and marketing, to understand which products are commonly bought in combination with one another. This information can be used, for example, as a starting point for recommendation systems (such as the notifications that “Customers who bought item X also bought item Y” which appear on the Amazon® website).

This report investigates the application of association rule analysis to the subjects that students choose to study at A and AS level, using data from the National Pupil Database (NPD).

Findings

In general the analysis revealed relatively little new information that was not already known, or obvious, but did serve as an illustration of the technique and the types of relationships uncovered.

Particular results of interest from the analysis of student subject choice were:

- The strongest associations were found between Science subjects (including Mathematics). A student was much more likely to be taking a Science subject if he/she was already studying another one, and the numbers of students to whom these rules applied (the support) were large.
- There were also strong associations between classical subjects, although the numbers of students taking these subjects was much lower. For example, a student studying Latin was much more likely than average to also be studying Ancient Greek (but still unlikely on an absolute level).
- Some differences in uptake of subjects across subgroups appeared to be more dramatic when considered as part of combinations rather than at a single-subject level. For example, the gender differences in uptake of Science subjects.

Recommendations

- The use of association rules is best suited to initial exploration of unfamiliar data, to enable hypotheses to be formed that can be investigated using other methods.

1 Introduction

1.1 Context

At age 16, students in England face a wide variety of choices. The most popular pathway is to continue academic study via AS and A levels. Under the current system, most of these students take three or four subjects at AS level (Sutch, 2014), drop one subject, and then proceed to take full A levels in three subjects. Cambridge Assessment Research Division produces annual reports on the provision and uptake of A levels by subject (e.g., Gill, 2014b), which also present the most popular combinations of subjects.

It is of interest to look at combinations of subjects studied by pupils for several reasons. Firstly, to see the breadth, or indeed specialism, of the curriculum followed by students at this level. Secondly, students' understanding of and attainment in a particular subject may be influenced by whether they are studying related subjects. For example, if a student is struggling to understand a particular concept, covering a related concept or topic in a different way in another subject may help. Finally, university courses often have several prerequisite A levels, not just in the intended subject of study, and so admissions decisions are often influenced by the range of A levels taken by applicants. This latter point has come to the forefront in recent years, with the Russell Group publishing guidance to 16-year-olds on subject choice (Russell Group, 2011).

This report investigates the use of a data mining technique, association rules, in order to explore what choices students are making.

1.2 Association rules

Association rule analysis is a popular data mining method which finds values which occur frequently together. Its origins are in marketing, and understanding products that are commonly bought together. When applied to binary-valued data (that is, a product is either bought by a customer, or it is not) in this way, this technique is often referred to as *market basket analysis*.

A classic example of this is the supposed finding that beer and nappies were frequently bought together at a certain supermarket, the inference being that these customers were young fathers. This behaviour could then be exploited by the shop by, for example, placing the items near each other or increasing the price of beer. Although this example is pervasive, Power (2002) investigated its history and found that it had been (at the very least) embellished over time. Association rules form a starting point for online recommendation systems. For example, the notifications that "Customers who bought item X also bought item Y" appearing on the Amazon website, although they are typically merged with data relevant to the individual logged-in customer.

1.2.1 Explanation

The basic form of an association rule is $X \Rightarrow Y$ (which is read as 'if X then Y'). However, it is important to note that no causal relationship is implied; rather, that if a transaction contains item X there is a good chance it also contains item Y. The left hand side (LHS) of the rule is referred to as the antecedent, and the right hand side (RHS) as the consequent.

Reflecting its origins in retail and marketing, the key concepts are referred to as items[1] and transactions. The data is considered as a set of transactions, each containing one or more items such as beer or nappies. Although the terminology reflects the origins in retail and marketing, there is a fairly clear mapping into the domain of subject choice, as shown in Table 1:

[1] Not to be confused with items in a test.

Table 1: mapping to domain of subject choice

Term		*Subject domain*
Items	↦	Subjects
Transactions	↦	Students

More formally, using the definition expressed by García, Romero, Ventura, de Castro, and Calders (2009), we consider $I = i_1, \ldots, i_m$, a set of items, and D a set of transactions where each transaction T is a set of items such that $T \subseteq I$ (T is a subset of I). A transaction T *contains* X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$, and $X \cap Y = \emptyset$ (X and Y are distinct, with no items in common).

Despite the name, the rules are not absolute (that is, not *all* customers buying nappies also buy beer), and hold to varying degrees. A key piece of information is thus the *confidence* of a rule, which can be considered as the degree to which it holds: the rule $X \Rightarrow Y$ holds with confidence c if $c\%$ of transactions in D that contain X also contain Y.

In addition, rules may hold with high confidence but only in a small proportion of transactions. For example, two low-frequency items, which are rarely bought, but when they *are* bought almost always appear together. The *support* of the rule $X \Rightarrow Y$ is the proportion of all transactions that contain $X \cup Y$ (that is, the items in X and the items in Y).

When rules are mined, minimum support and confidence thresholds can be specified, but there are still typically a large number of rules generated. In order to sort through these, it is necessary to consider measures of so-called 'interestingness'. Geng and Hamilton (2006) set out nine criteria which are aspects to interestingness, some of which can be derived from data and probability theory alone, while other subjective or semantic measures require domain knowledge and user interaction. Several objective measures are available (Tan, Kumar, & Srivastava (2004) reviewed 21 such measures) but following recommendations by Merceron and Yacef (2008) and Luna Bazaldua, Baker, and San Pedro (2014) two measures are presented here. The first, and simpler of the two, is *lift* (referred to by Tan et al. (2004) as interest), which measures deviation from statistical independence, and thus the extent to which the rule outperforms what would be expected anyway: if the confidence of rule $X \Rightarrow Y$ is 0.5, but half of *all* transactions in D contain Y, the rule does not tell us anything new and lift is equal to 1. It is defined as:

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{confidence}(X \Rightarrow Y)}{\mathrm{support}(Y)} = \frac{\mathrm{support}(X \Rightarrow Y)}{\mathrm{support}(X)\mathrm{support}(Y)}$$

(the number of transactions containing both X and Y, divided by the product of the number containing X and the number containing Y).

The second measure used is *cosine*. This is defined similarly, but adds more emphasis to rules with higher support. Unlike lift, cosine has the attractive property that transactions containing neither X nor Y do not influence the measure[2] (this is known as null-invariance).

$$\mathrm{cosine}(X \Rightarrow Y) = \frac{\mathrm{support}(X \Rightarrow Y)}{\sqrt{\mathrm{support}(X)\mathrm{support}(Y)}}$$

[2] Consider 100 transactions, of which 20 contain item X, 30 contain item Y and 15 contain both items X and Y. Then $\mathrm{lift}(X \Rightarrow Y) = 0.15/(0.2 \times 0.3) = 2.5$, and $\mathrm{cosine}(X \Rightarrow Y) = 0.15/\sqrt{0.2 \times 0.3} = 0.612$. If the number of transactions increases to 200, but with the number containing X, Y and both X and Y remaining the same, $\mathrm{lift}(X \Rightarrow Y)$ increases to $0.075/(0.1 \times 0.15) = 5$, whereas $\mathrm{cosine}(X \Rightarrow Y)$ remains 0.612.

Merceron and Yacef (2008) state that for strong symmetric association rules, where the support of X, Y and $X \Rightarrow Y$ are large, cosine performs better than lift. They recommended (writing in an e-learning context) that rules with a cosine of around 0.65 or lower are rejected.

In a study, considering rules generated from e-learning systems, in which expert ratings of interestingness were compared to measures calculated from the data, Luna Bazaldua et al. (2014) found that lift and cosine were both important, but interestingly when support and confidence were accounted for, cosine was *negatively* correlated with interestingness. The authors suggested that this might indicate the value in rarity, turning up rules that were surprising to the experts.

Filtering the rules manually depending on the results is a common technique for reducing the number of rules: For example, constraining the rules so that the RHScontains, or does not contain, a particular item. This is effectively using subjective and semantic measures of interestingness to arrive at the most useful set of rules.

Association rules are a fairly simple concept, but provide a useful framework for analysis. The underlying data need not be binary only: categorical variables can easily be added through means of dummy variables, and continuous variables can be recoded to discrete categories, based on quantiles for example. Importantly, a number of efficient algorithms have been developed in order to generate association rules. The most popular is *Apriori* (Agrawal, Imielinski, & Swami, 1993), which works on the principle that if an itemset is frequent, all of its subsets are frequent. That is, if the combination {Physics, Chemistry, Biology} is frequent then so must be {Physics, Chemistry}, {Chemistry, Biology}, {Physics, Biology} as well as each of the three subjects individually. As a consequence, if an itemset is *infrequent* (that is, its support is lower than a pre-determined threshold) then its *supersets* are also infrequent so need not be considered further.

Further research and development has been carried out on extending association rules. One important area is looking for contrasting sets of the data (Bay & Pazzani, 2001) and discovering interesting subgroups. Kralj Novak, Lavřac, and Webb (2009) present a unifying survey of this area.

1.2.2 Application in educational research

Over recent years, an educational data mining community has emerged[3]. Researchers have used a variety of techniques, but most applications have been concentrated on the data-rich field of e-learning, in order to understand how students interact with these systems (at a detailed level) and what behaviours are associated with a better mark. There have been several 'state of the art' articles (e.g., Romero & Ventura, 2010) but the most comprehensive overview currently is by Romero, Ventura, Pechenizkiy, and Baker (2011).

García, Romero, Ventura, and de Castro (2011) describe the use of the *Predictive Apriori algorithm* (a variant of Apriori where the number of rules to be mined is pre-determined) to generate association rules about e-learning courses (using data at a student level) which were then filtered by experts, formulated as 'problems' and 'recommendations'. These were then presented to teachers who went on to use them to improve courses.

The activity of assessment generates large quantities of data, and this is ideally suited to data mining tasks. Kumar and Chadha (2012) applied association rules to data on student grades in a Computing course at an Indian university, and derived rules which found (unsurprisingly) relationships between performance in different assessments on the same subject. They also found that a general lack of progress in certain subjects suggested deeper issues.

[3]For more information, see the International Educational Data Mining Society homepage: http://educationaldatamining.org/

Romero, Romero, Luna, and Ventura (2010) have applied another variant of the Apriori algorithm which focuses on rare rules, that is rules with low support but still high confidence, to e-learning data. Such rules can be useful in an educational setting because some rare events, such as student drop-out or failure, are particularly interesting to study. A different algorithm is necessary to do this because simply setting low minimum support thresholds can result in an unmanageable number of patterns, with most of them being frequent.

Minaei-Bidgoli, Tan, and Punch (2004) presented a general formulation for contrast rules, using different measures of interestingness, and applied this to data generated by an e-learning system. The authors used Apriori to generate rules with low minimum support, and then found the common rules between two contrast subsets (which were created based on values of a binary variable). The measures employed were difference in confidence of the contrast rules $\{X \Rightarrow Y\}$ and $\{X \Rightarrow \overline{Y}\}$ (where $\overline{Y}$ indicates transactions not containing Y)[4], difference of proportions, comparing the confidence of $\{X \Rightarrow Y\}$ and $\{\overline{X} \Rightarrow Y\}$, and chi-square.

1.2.3 Other data mining techniques on similar data

Vialardi, Bravo, Shafti, and Ortigosa (2009) describe the application of similar techniques in order to recommend individual Higher Education (HE) modules to students on a Computer Science course. The *C4.5 classifier algorithm*, rather than association rules, was used to generate a decision tree.

Singleton (2009) used data on applicants' UCAS course choices in the 2004 application cycle to explore how strong their subject preferences were (that is, how consistent they were across the courses applied to), with the intention of developing a decision support tool for HE institutions. Association rules were not formally used but there is undoubtedly an overlap with his method.

Another development in the educational data mining community is curriculum mining (Pechenizkiy, Trčka, De Bra, & Toledo, 2012). This is an application of process mining on the curriculum followed (that is, the courses chosen) by students at a university. The authors have used curriculum mining to uncover popular paths, test students' behaviour against constraints (for example, whether pre-requisites for particular courses have actually been passed), link the path through the system to student attainment, and model 'what-if' scenarios.

2 Aim of this research

This report aims to answer the following question:

- Are association rules a useful tool for investigating A level subject choice and uptake at a student level?

3 Data/methods

The data for this analysis come from the National Pupil Database (NPD), maintained by the Department for Education. Specifically, the dataset extracted for the recent Statistics Report on uptake of AS levels (Sutch, 2014) was used; this covers students who were in Year 13 in 2012, and therefore typically took GCSEs in 2010, AS units in 2011 and A2 units in 2012. It includes the subjects taken by each student at AS or A level. The following information was extracted:

- Gender
- Mean Key Stage 4 (KS4) score (typically GCSE but also including IGCSE), condensed to terciles (low, medium and high)

[4] Because Y and $\overline{Y}$ are mutually exclusive, $\text{confidence}(X \Rightarrow Y) + \text{confidence}(X \Rightarrow \overline{Y}) \equiv 1$, so finding rules with the highest difference in confidence using this definition is equivalent to simply finding the rules with highest confidence of $\{X \Rightarrow Y\}$.

- Centre type (Academy, Comprehensive, Further Education (FE) College, Grammar, Independent, Secondary Modern, Sixth Form College, Tertiary College, Other).

Note that this dataset does not include students taking other 16–18 qualifications such as the Cambridge Pre-U.

The total number of students was 294,458, with 97 distinct subjects. The most frequent subjects were Mathematics, Psychology, Biology, General Studies and Chemistry.[5]

The arules package (Hahsler, Buchta, Gruen, & Hornik, 2014) for R (R Core Team, 2014) was used to generate the association rules using the Apriori algorithm. Low minimum values for support and confidence were chosen (support of 0.0001, corresponding to 30 students or more, and a confidence of 0.001), in order to generate a large number of rules which could be investigated and potentially filtered out later. Rules of length 1 (that is, where the antecedent was the empty set) were excluded, as these simply reflect the uptake of each subject individually (as shown in Sutch (2014) for example).

It is important to note that the itemsets do not necessarily constitute the whole of a student's A level choices. For example, the support for the rule {French ⇒ German} refers to students with *at least* French and German; that is, students with {English Literature, French, German}, {French, German} and {French, German, Mathematics} would all be included. This is different to the combinations regularly presented in Statistics Reports, for example Sutch (2014).

A number of different sets of rules were generated:

1. Basic rules, considering only the subjects chosen by each student, and thus representing a pure market basket analysis. An example transaction would consist of:

   {Mathematics
   Psychology
   Law}

2. Rules from transactions amended to include background information on the student. An example transaction might consist of:

   {subject=Mathematics
   subject=Psychology
   subject=Law
   schooltype=Comprehensive
   gender=F
   mean_gcse=Low}

3. Basic rules, as with 1, but generated using subsets of students (for example, female students; students at comprehensive schools). As a result, the support values generated are as a proportion of only the subset. This allows easy comparison of rules across the subpopulations, using a similar idea as Minaei-Bidgoli et al. (2004) and reporting differences of proportions.[6]

[5] Full statistics for uptake of each subject individually are available in Sutch (2014).

[6] Implementation of a more sophisticated algorithm for generating contrast sets such as STUCCO (Bay & Pazzani, 2001) was considered, but no such algorithms were included in arules, nor available elsewhere in an easily transferable form, so this would have required substantial low-level programming.

4 Results

4.1 Results: basic rules

The figure below presents the values of the Jaccard distance between pairs of subjects. The Jaccard distance is the complement of the ratio of the intersection to the union of two sets; that is, the complement of the number of students who took *both* subjects as a proportion of the students who took *either* subject. As such it is symmetric. In terms of support it is as follows:

$$d_J(X,Y) = 1 - \frac{\text{support(X,Y)}}{\text{support}(X) + \text{support}(Y) - \text{support}(X,Y)}$$

The darker cells in the figure indicate pairs of subjects which are commonly taken together. These include the Science subjects (for example Mathematics and Physics) as well as other more isolated pairs, such as Latin and Classical Greek.

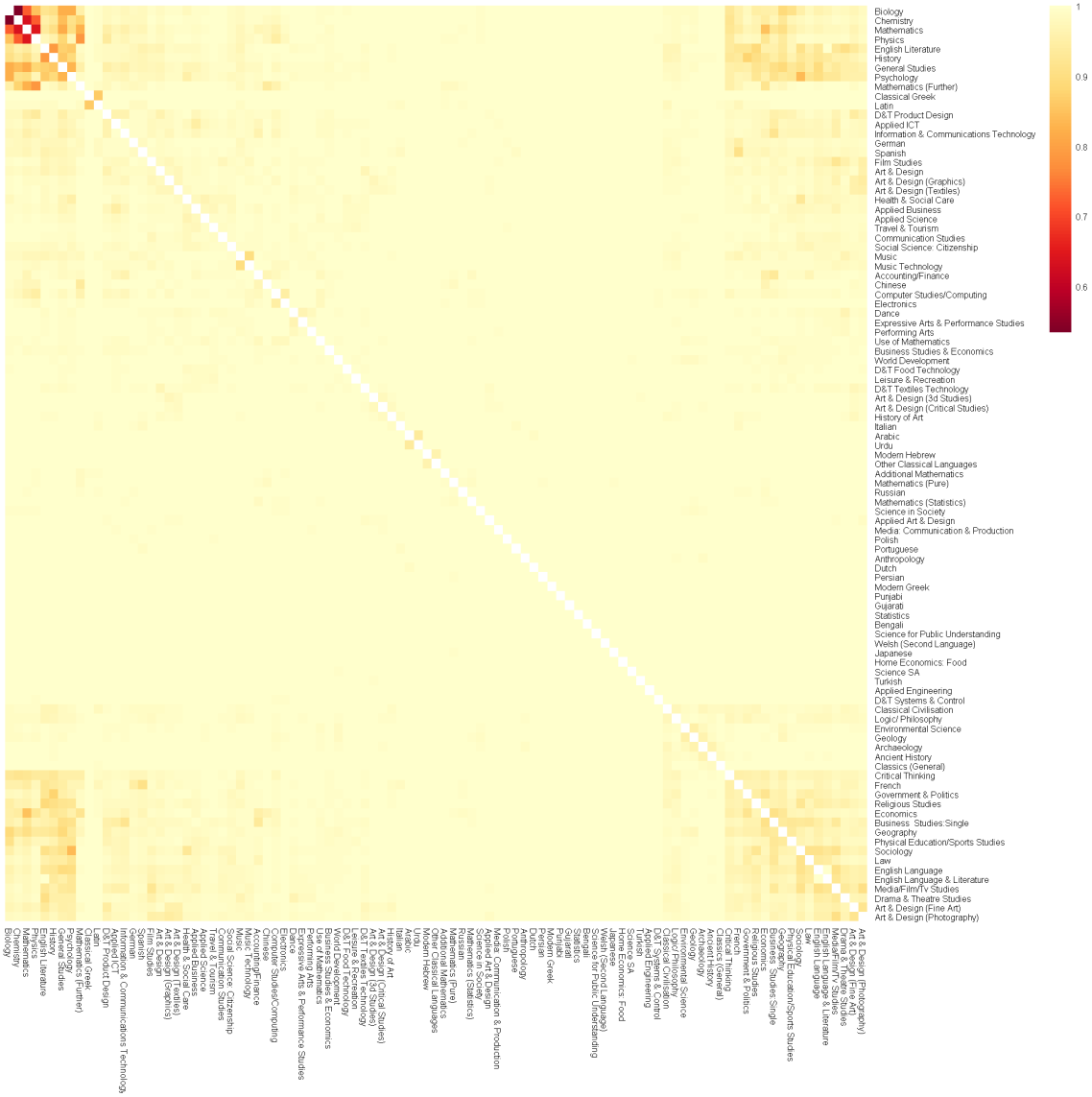

Figure 1: Heatmap of Jaccard distance between pairs of subjects

The Jaccard distance between pairs of subjects has been used to perform a cluster analysis, using Ward's method, and the resulting dendrogram is shown in Figure 2. The clustering identifies the major science subjects (Biology, Chemistry, Physics, Mathematics and Further

Mathematics) as a distinct cluster, but other than these there are few obvious clusters involving more than two subjects. The next most distinct groups are the common European languages (French, German and Spanish), and a Humanities cluster (Government & Politics, History, English Literature, and Religious Studies).

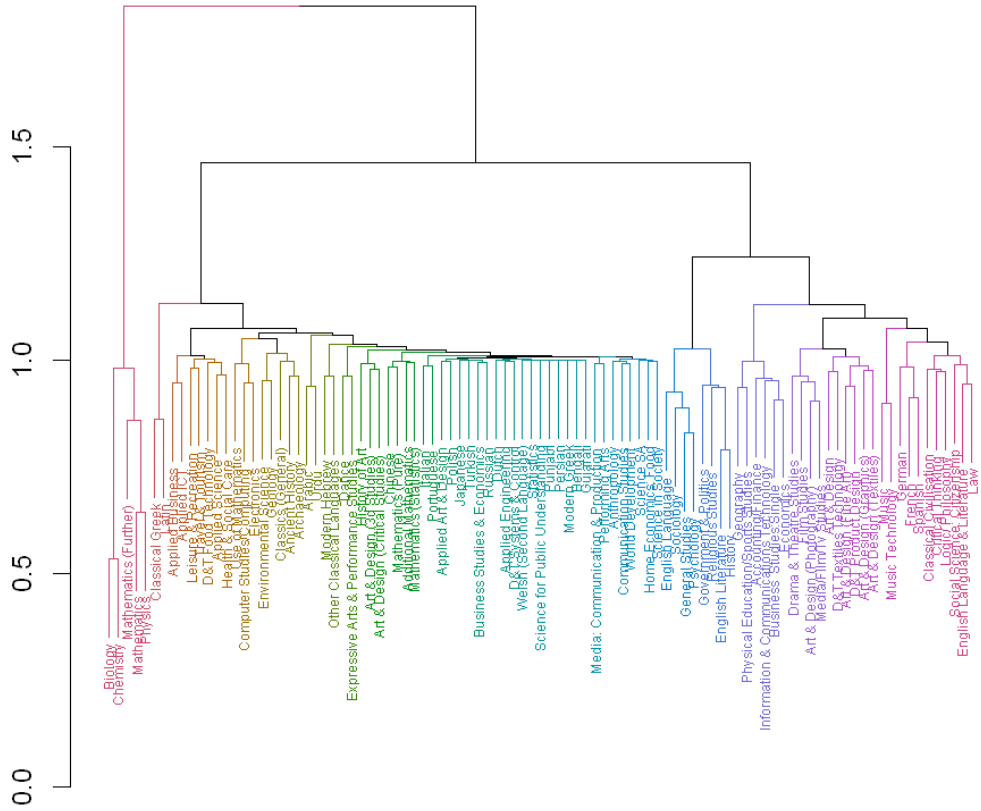

Figure 2: Dendrogram from hierarchical clustering of subjects based on Jaccard distance

The Apriori algorithm generated 47,677 rules with minimum support 0.0001 and confidence 0.001. Figure 3 shows the relationship between support, confidence, lift and cosine for rules with support of at least 0.01.

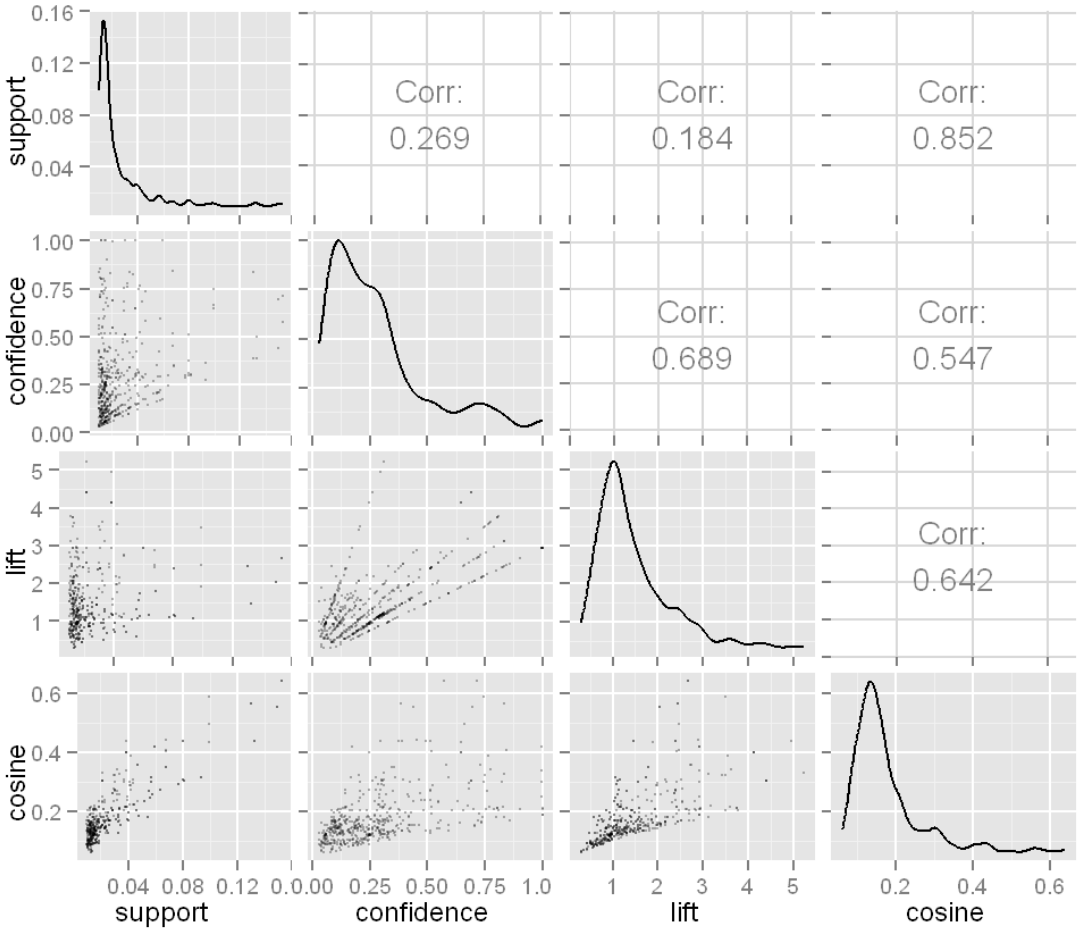

Figure 3: Characteristics of, and relationship between, interestingness measures of rules

There is a very high correlation between support and cosine, and slightly lower correlations between cosine and lift, and between confidence and lift. Most of the rules (under any measure) are not interesting, having low support, confidence and cosine, and a lift near to 1. The top 10 rules ordered by each of support, confidence, lift and cosine are presented in Tables 2–5 below.

Table 2: Top 10 support

rules	*support*	*confidence*	*lift*	*cosine*
{Chemistry} ⇒ {Biology}	0.1532	0.7142	2.670	0.6395
{Biology} ⇒ {Chemistry}	0.1532	0.5727	2.670	0.6395
{Chemistry} ⇒ {Mathematics}	0.1494	0.6967	2.043	0.5525
{Mathematics} ⇒ {Chemistry}	0.1494	0.4382	2.043	0.5525
{Biology} ⇒ {Mathematics}	0.1326	0.4957	1.453	0.4390
{Mathematics} ⇒ {Biology}	0.1326	0.3888	1.453	0.4390
{Physics} ⇒ {Mathematics}	0.1306	0.8337	2.444	0.5651
{Mathematics} ⇒ {Physics}	0.1306	0.3830	2.444	0.5651
{Biology, Mathematics} ⇒ {Chemistry}	0.0991	0.7474	3.484	0.5876
{Chemistry, Mathematics} ⇒ {Biology}	0.0991	0.6632	2.479	0.4957

Table 3: Top 10 confidence

rules	*support*	*confidence*	*lift*	*cosine*
{General Studies, Mathematics (Further)} ⇒ {Mathematics}	0.0150	1	2.932	0.2099
{General Studies, Mathematics (Further),Physics} ⇒ {Mathematics}	0.0099	1	2.932	0.1705
{Chemistry, General Studies, Mathematics (Further)} ⇒ {Mathematics}	0.0079	1	2.932	0.1520
{History, Mathematics (Further)} ⇒ {Mathematics}	0.0057	1	2.932	0.1293
{Economics, Mathematics (Further), Physics} ⇒ {Mathematics}	0.0057	1	2.932	0.1287
{Critical Thinking, Mathematics (Further)} ⇒ {Mathematics}	0.0056	1	2.932	0.1278
{Chemistry, General Studies, Mathematics (Further), Physics} ⇒ {Mathematics}	0.0054	1	2.932	0.1257
{Biology, General Studies, Mathematics (Further)} ⇒ {Mathematics}	0.0040	1	2.932	0.1077
{Critical Thinking, Mathematics (Further), Physics} ⇒ {Mathematics}	0.0037	1	2.932	0.1043
{French, Mathematics (Further)} ⇒ {Mathematics}	0.0036	1	2.932	0.1034

Table 4: Top 10 lift

rules	*support*	*confidence*	*lift*	*cosine*
{Classical Greek, History} ⇒ {Latin}	4e-04	0.7836	130.4	0.2157
{Critical Thinking, Latin} ⇒ {Classical Greek}	1e-04	0.1606	128.2	0.1162
{Classical Greek, History, Mathematics} ⇒ {Latin}	1e-04	0.7500	124.8	0.1235
{History, Latin ,Mathematics} ⇒ {Classical Greek}	1e-04	0.1552	123.8	0.1230
{Classical Greek, French} ⇒ {Latin}	2e-04	0.7246	120.6	0.1431
{Classical Greek, Critical Thinking} ⇒ {Latin}	1e-04	0.7209	120.0	0.1124
{History, Latin} ⇒ {Classical Greek}	4e-04	0.1496	119.4	0.2063
{Latin, Mathematics} ⇒ {Classical Greek}	4e-04	0.1480	118.1	0.2230
{Latin, Mathematics (Further)} ⇒ {Classical Greek}	1e-04	0.1473	117.5	0.1232
{Latin, Mathematics, Mathematics (Further)} ⇒ {Classical Greek}	1e-04	0.1473	117.5	0.1232

Table 5: Top 10 cosine

rules	*support*	*confidence*	*lift*	*cosine*
{Chemistry} ⇒ {Biology}	0.1532	0.7142	2.670	0.6395
{Biology} ⇒ {Chemistry}	0.1532	0.5727	2.670	0.6395
{Biology, Mathematics} ⇒ {Chemistry}	0.0991	0.7474	3.484	0.5876
{Physics} ⇒ {Mathematics}	0.1306	0.8337	2.444	0.5651
{Mathematics} ⇒ {Physics}	0.1306	0.3830	2.444	0.5651
{Chemistry} ⇒ {Mathematics}	0.1494	0.6967	2.043	0.5525
{Mathematics} ⇒ {Chemistry}	0.1494	0.4382	2.043	0.5525
{Chemistry, Mathematics} ⇒ {Biology}	0.0991	0.6632	2.479	0.4957
{Chemistry, Mathematics} ⇒ {Physics}	0.0675	0.4517	2.883	0.4411
{Mathematics} ⇒ {Biology}	0.1326	0.3888	1.453	0.4390

Each of these lists shows different aspects:

The rules with highest support (Table 2) show the relationships between the various science subjects, which feature in the most popular subject combinations (Sutch, 2014, Section 4). They appear in pairs because the support of $X \Rightarrow Y$ is the same as the support of $Y \Rightarrow X$ (the proportion of students taking both X and Y). These have relatively low lift (although greater than 1) but a variety of confidences: the largest is {Physics} ⇒ {Mathematics}, indicating that 83 per cent of students taking Physics also took Mathematics. The confidences for the rules where Mathematics is the consequent right hand side are higher than those for which it is the antecedent left hand side, because it is the most popular subject anyway. The lift of Biology ⇒ Mathematics and vice versa is lower because both of these subjects are popular.

The rules with highest confidence (Table 3) all have confidence 1, indicating that the rule held in all relevant cases (for example, all candidates with History and Further Mathematics also took Mathematics). However, crucially, the support of all these rules is fairly low: the highest is 0.0150. The clear dependency of Further Mathematics on Mathematics is not surprising and does not tell us anything of interest.

The rules with highest lift (Table 4) all relate to classical languages. These subjects have low uptake (and as such, support is low), but are (relatively) commonly taken together - far more likely than one would expect by chance. Although only 105 students (0.00036, as a proportion) took Classical Greek, History and Latin as part of their AS choices, 78 per cent of those taking Classical Greek and History also took Latin - 130 times more likely than would be expected by chance.

When inspecting the rules with highest cosine (Table 5), the first thing to note is that the highest value is only 0.64, lower than the threshold recommended by Merceron and Yacef (2008) of 0.65. However, the nature of the data is rather different and their advice may be less applicable to this context. There is a large overlap with the rules with the highest cosine and those with the highest support, so the cosine seems not to be contributing much in addition to support here.

In order to reduce the number of rules further and produce a more ‘interesting’ subset, rules which included Mathematics or Further Mathematics on either side were removed, and support was filtered to 0.01 (representing 2,945 students).

Table 6: Top 10 lift for rules with support > 0.01, excluding Mathematics on either side

rules	*support*	*confidence*	*lift*	*cosine*
{Biology, General Studies, Physics} ⇒ {Chemistry}	0.0127	0.7695	3.587	0.2131
{Biology, Physics} ⇒ {Chemistry}	0.0421	0.7623	3.554	0.3866
{Biology, Critical Thinking} ⇒ {Chemistry}	0.0151	0.6783	3.162	0.2188
{Chemistry, Psychology} ⇒ {Biology}	0.0339	0.8335	3.115	0.3251
{Chemistry, General Studies, Psychology} ⇒ {Biology}	0.0104	0.8335	3.115	0.1800
{Chemistry, Critical Thinking} ⇒ {Biology}	0.0151	0.7530	2.815	0.2064
{English Literature, Government & Politics} ⇒ {History}	0.0100	0.5562	2.805	0.1676
{Chemistry, General Studies} ⇒ {Biology}	0.0467	0.7326	2.738	0.3576
{Biology, General Studies} ⇒ {Chemistry}	0.0467	0.5817	2.712	0.3558
{Chemistry, English Literature} ⇒ {Biology}	0.0133	0.7234	2.704	0.1900

Table 7: Top 10 cosine for rules with support > 0.01, excluding Mathematics on either side

rules	*support*	*confidence*	*lift*	*cosine*
{Chemistry} ⇒ {Biology}	0.1532	0.7142	2.670	0.6395
{Biology} ⇒ {Chemistry}	0.1532	0.5727	2.670	0.6395
{Physics} ⇒ {Chemistry}	0.0802	0.5118	2.386	0.4374
{Chemistry} ⇒ {Physics}	0.0802	0.3739	2.386	0.4374
{Biology, Physics} ⇒ {Chemistry}	0.0421	0.7623	3.554	0.3866
{Chemistry, General Studies} ⇒ {Biology}	0.0467	0.7326	2.738	0.3576
{Biology, General Studies} ⇒ {Chemistry}	0.0467	0.5817	2.712	0.3558
{English Literature} ⇒ {History}	0.0681	0.3527	1.778	0.3481
{History} ⇒ {English Literature}	0.0681	0.3435	1.778	0.3481
{Chemistry, Psychology} ⇒ {Biology}	0.0339	0.8335	3.115	0.3251

The tables each feature some rules relating History to English Literature, but the other rules relate to Science subjects (even though Mathematics had been specifically excluded).

Perhaps the most interesting insights are gleaned by comparing rules from the tables. For example, the confidence of Biology ⇒ Chemistry is 0.5727 (lift 2.670), but when Physics is also included, the confidence of {Biology, Physics} ⇒ {Chemistry} is 0.7623 (lift 3.554). That is, students taking two Sciences are likely to be taking a third. Similarly, candidates with Chemistry are more likely to study Biology if they are also taking Psychology (confidence 0.8335 with Psychology compared to 0.7142 without), and this seems reasonable: Biology seems notionally to sit between Psychology and Chemistry as there are linkages with both.

The following tables show the rules with highest lift and cosine which do not include any Science subjects (that is, Mathematics, Biology, Chemistry or Physics) on the right hand side, and have support of at least 0.01.

Table 8: Top 10 lift for rules with no Science on RHS

rules	*support*	*confidence*	*lift*	*cosine*
{English Literature, Government & Politics} ⇒ {History}	0.0100	0.5562	2.805	0.1676
{Health & Social Care} ⇒ {Sociology}	0.0112	0.3286	2.436	0.1652
{Sociology} ⇒ {Health & Social Care}	0.0112	0.0831	2.436	0.1652
{English Literature, History} ⇒ {Government & Politics}	0.0100	0.1469	2.398	0.1549
{Government & Politics} ⇒ {History}	0.0287	0.4689	2.365	0.2607
{History} ⇒ {Government & Politics}	0.0287	0.1449	2.365	0.2607
{Art & Design (Photography)} ⇒ {Media/Film/TV Studies}	0.0146	0.2345	2.215	0.1796
{Media/Film/TV Studies} ⇒ {Art & Design (Photography)}	0.0146	0.1375	2.215	0.1796
{History, Religious Studies} ⇒ {English Literature}	0.0105	0.4209	2.178	0.1509
{Information & Communications Technology} ⇒ {Business Studies: Single}	0.0126	0.2645	2.166	0.1650

Table 9: Top 10 cosine for rules with no Science on RHS

rules	*support*	*confidence*	*lift*	*cosine*
{English Literature} ⇒ {History}	0.0681	0.3527	1.778	0.3481
{History} ⇒ {English Literature}	0.0681	0.3435	1.778	0.3481
{Mathematics} ⇒ {Economics}	0.0574	0.1684	1.732	0.3154
{Mathematics} ⇒ {General Studies}	0.0935	0.2742	1.061	0.3150
{Biology} ⇒ {General Studies}	0.0803	0.3000	1.161	0.3053
{Biology} ⇒ {Psychology}	0.0815	0.3045	1.120	0.3021
{Sociology} ⇒ {Psychology}	0.0569	0.4219	1.552	0.2972
{Psychology} ⇒ {Sociology}	0.0569	0.2093	1.552	0.2972
{Psychology} ⇒ {General Studies}	0.0772	0.2840	1.099	0.2913
{General Studies} ⇒ {Psychology}	0.0772	0.2988	1.099	0.2913

The rules with highest lift feature a number of reciprocal links. For example, the link between History and Government & Politics, and the link between Sociology and Health & Social Care. The confidence for some of these rules is rather low, but this reflects the low uptake of these subjects anyway. The rules with highest cosine, on the other hand, highlight common pairings, such as History and English Literature, but do not probe deeper. Once again, these rules show that support has rather too much influence on the cosine, for example Mathematics ⇒ General Studies, which has reasonable confidence but a lift very close to 1.

4.2 Results: demographic information included in transactions

When background information was included in the transactions, the Apriori algorithm generated 707,502 rules with minimum support and confidence 0.0001 and 0.001 respectively. The support, confidence, lift and cosine are plotted against each other in Figure 4for rules with support at least 0.01.

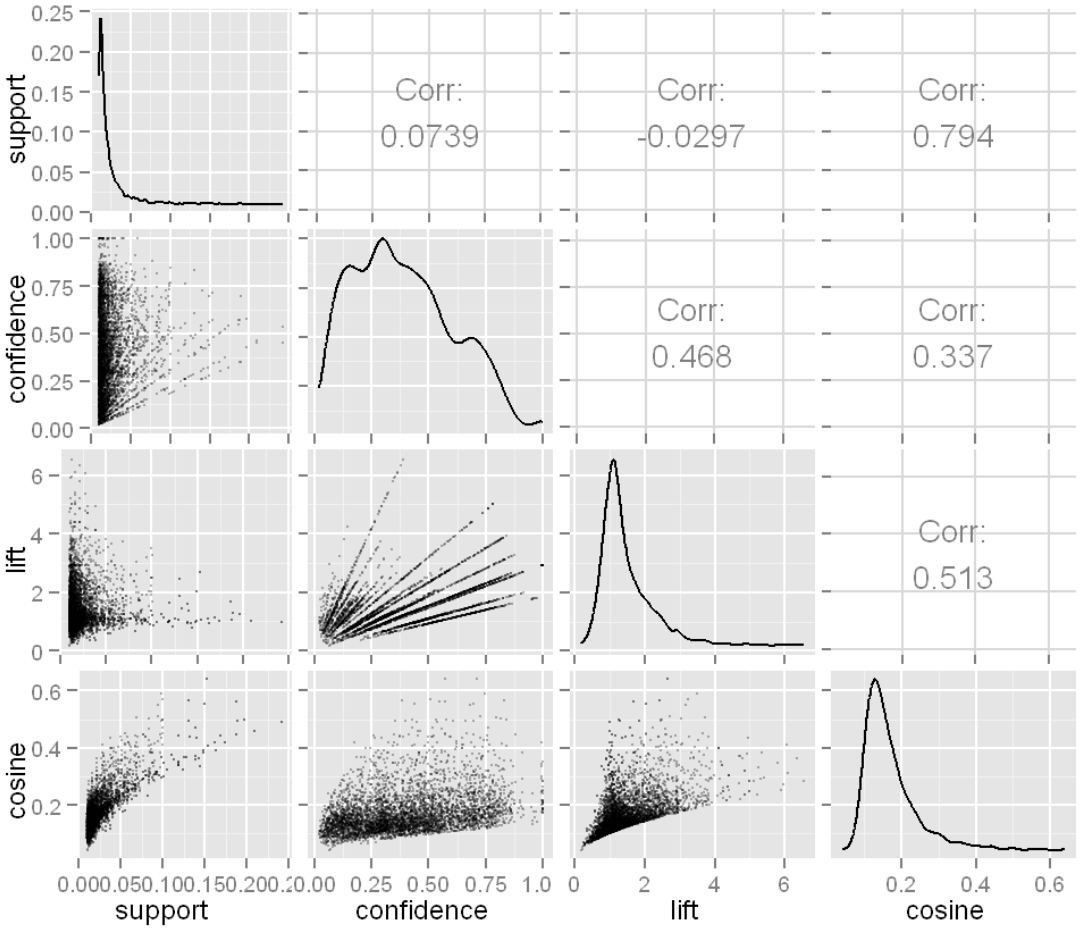

Figure 4: Characteristics of, and relationship between, interestingness measures

The rules generated in the previous section (which relate to subjects only) still apply to the enhanced dataset, so our objective must be to find new rules which include the demographic variables. We therefore need to filter based on certain patterns of rules, as well as numerical interestingness measures.

Although we are most interested in patterns of subject uptake, first we examine rules that contain no subject information in either the left or right hand side, in order to see the relationships between the background variables. The top ten rules by confidence, lift and cosine are presented below.

The rules with highest lift and confidence overlap substantially, and contain many rules showing a relationship between good GCSE results and attendance at a Grammar or Independent school. Conversely, students attending a Secondary Modern school are more likely than average to be in the lowest group in terms of mean GCSE. These rules are not surprising but are useful to bear in mind when considering other results.

The rules with highest cosine are somewhat different. While their support is high, their lift values are close to 1 and thus the presence or absence of the LHS of the rule has little bearing on the right hand side. For example, the rule with highest cosine is {Comprehensive school} ⇒ {Female student}, but the confidence is only 0.5371 (slightly lower than the proportion of female students in the dataset, 0.5392). Too much value is being placed on support.

Table 10: Rules with no subject, by confidence

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
GENDER=F, schooltype=Grammar	mean_gcse=High	0.0285	0.6931	2.122	0.2461
schooltype=Grammar	mean_gcse=High	0.0513	0.6592	2.019	0.3219
mean_gcse=High, schooltype=FE College	GENDER=F	0.0036	0.6453	1.197	0.0655
GENDER=F, schooltype=Independent	mean_gcse=High	0.0373	0.6378	1.953	0.2700
GENDER=M, schooltype=Grammar	mean_gcse=High	0.0228	0.6213	1.902	0.2083
mean_gcse=Med, schooltype=FE College	GENDER=F	0.0094	0.6204	1.151	0.1041
GENDER=M, schooltype=Secondary Modern	mean_gcse=Low	0.0050	0.6165	1.900	0.0972
mean_gcse=High, schooltype=Tertiary College	GENDER=F	0.0061	0.6138	1.138	0.0832
schooltype=Independent	mean_gcse=High	0.0720	0.6124	1.875	0.3674
mean_gcse=Low, schooltype=Independent	GENDER=M	0.0056	0.6080	1.319	0.0857

Table 11: Rules with no subject, by lift

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
GENDER=F ,schooltype=Grammar	mean_gcse=High	0.0285	0.6931	2.122	0.2461
GENDER=M, mean_gcse=High	schooltype=Independent	0.0347	0.2490	2.118	0.2710
GENDER=M, mean_gcse=High	schooltype=Grammar	0.0228	0.1638	2.103	0.2190
schooltype=Grammar	mean_gcse=High	0.0513	0.6592	2.019	0.3219
mean_gcse=High	schooltype=Grammar	0.0513	0.1572	2.019	0.3219
GENDER=F, mean_gcse=High	schooltype=Grammar	0.0285	0.1523	1.956	0.2362
GENDER=F, schooltype=Independent	mean_gcse=High	0.0373	0.6378	1.953	0.2700
GENDER=M, schooltype=Grammar	mean_gcse=High	0.0228	0.6213	1.902	0.2083
GENDER=M, schooltype=Secondary Modern	mean_gcse=Low	0.0050	0.6165	1.900	0.0972
schooltype=Independent	mean_gcse=High	0.0720	0.6124	1.875	0.3674

Table 12: Rules with no subject, by cosine

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
schooltype=Comprehensive	GENDER=F	0.2424	0.5371	0.9961	0.4913
GENDER=F	schooltype=Comprehensive	0.2424	0.4495	0.9961	0.4913
schooltype=Comprehensive	GENDER=M	0.2089	0.4629	1.0045	0.4581
GENDER=M	schooltype=Comprehensive	0.2089	0.4533	1.0045	0.4581
GENDER=F	mean_gcse=High	0.1873	0.3474	1.0638	0.4464
mean_gcse=High	GENDER=F	0.1873	0.5736	1.0638	0.4464
mean_gcse=Low	schooltype=Comprehensive	0.1690	0.5206	1.1538	0.4415
schooltype=Comprehensive	mean_gcse=Low	0.1690	0.3745	1.1538	0.4415
mean_gcse=Low	GENDER=M	0.1614	0.4973	1.0791	0.4173
GENDER=M	mean_gcse=Low	0.1614	0.3502	1.0791	0.4173

Having considered the background variables, we now turn our attention to rules where the right hand side (consequent) is a subject, and the left hand side contains gender, school type, and mean GCSE group in turn. Considering gender first of all, the rules with the highest lift are similar to those already discovered (in the previous section). The consequents of the rules are all Classical Greek, as with Table 4. The rule with highest lift, for example, {Male student, Independent school, top mean GCSE group, History, Latin} ⇒ {Classical Greek}, is very similar to one of the rules with highest lift (Table 4) with no additional variables – {History, Latin} ⇒ {Classical Greek} – but with lower support. Because the confidence of the rule, with the additional variables, is higher, the extra conditions are adding something here but it is not clear that this is the main determinant, or simply a result of random variation on such a small subset of data. The same applies to the other rules shown.

When cosine is used instead, a more diverse set of rules is obtained and this does seem to highlight some differences in the combinations, by comparing the confidences with those seen earlier in Section 4.1. For example, the probability that a male student taking Mathematics is also taking Physics is 0.5099, compared to the average for all students taking Mathematics of 0.3830 (Table 2). The set of rules also contains some basic examples, without a subject in the antecedent (for example {Male student} ⇒ {Mathematics}). For these rules, the lift is most interesting as it shows how different uptake of these subjects is among subgroups.

Table 13: Top 10 gender ⇒ subject by lift

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
GENDER=M,SUBJECT=History,SUBJECT=Latin ,mean_gcse=High,schooltype=Independent	SUBJECT=Classical Greek	2e-04	0.2137	170.5	0.1752
GENDER=M,SUBJECT=History, SUBJECT=Latin, schooltype=Independent	SUBJECT=Classical Greek	2e-04	0.2093	167.0	0.1750
GENDER=M,SUBJECT=Latin, SUBJECT=Mathematics, mean_gcse=High, schooltype=Independent	SUBJECT=Classical Greek	2e-04	0.2045	163.2	0.1869
GENDER=M,SUBJECT=Latin, mean_gcse=High, schooltype=Independent	SUBJECT=Classical Greek	4e-04	0.2037	162.6	0.2574
GENDER=M,SUBJECT=Latin, SUBJECT=Mathematics, schooltype=Independent	SUBJECT=Classical Greek	2e-04	0.2032	162.1	0.1877
GENDER=M,SUBJECT=Latin, schooltype=Independent	SUBJECT=Classical Greek	4e-04	0.2016	160.9	0.2593
GENDER=M,SUBJECT=English Literature, SUBJECT=Latin, schooltype=Independent	SUBJECT=Classical Greek	1e-04	0.1966	156.9	0.1366
GENDER=M,SUBJECT=English Literature, SUBJECT=Latin, mean_gcse=High, schooltype=Independent	SUBJECT=Classical Greek	1e-04	0.1954	155.9	0.1342
GENDER=M,SUBJECT=History, SUBJECT=Latin, mean_gcse=High	SUBJECT=Classical Greek	2e-04	0.1918	153.1	0.1781
GENDER=M,SUBJECT=History, SUBJECT=Latin	SUBJECT=Classical Greek	2e-04	0.1862	148.6	0.1769

Table 14: Top 10 gender ⇒ subject by cosine

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
GENDER=M,SUBJECT=Mathematics	SUBJECT=Physics	0.1004	0.5099	3.254	0.5715
GENDER=F,SUBJECT=Chemistry	SUBJECT=Biology	0.0831	0.8167	3.053	0.5036
GENDER=M,SUBJECT=Physics	SUBJECT=Mathematics	0.1004	0.8437	2.474	0.4983
GENDER=M	SUBJECT=Mathematics	0.1969	0.4273	1.253	0.4967
GENDER=F	SUBJECT=Psychology	0.1893	0.3512	1.292	0.4946
GENDER=F,SUBJECT=Biology	SUBJECT=Chemistry	0.0831	0.5575	2.599	0.4647
GENDER=M,SUBJECT=Mathematics, mean_gcse=High	SUBJECT=Physics	0.0554	0.5717	3.648	0.4497
GENDER=M, mean_gcse=High	SUBJECT=Mathematics	0.0970	0.6965	2.042	0.4450
GENDER=M	SUBJECT=Physics	0.1190	0.2582	1.648	0.4428
GENDER=M,SUBJECT=Biology	SUBJECT=Chemistry	0.0701	0.5917	2.758	0.4398

Similarly to the gender results, all but one rule with high lift for centre type and mean GCSE subgroups concerned the uptake of Classical Greek and therefore only the results for cosine are presented here (Tables 15 and 16). For centre type, the values of cosine are lower than for subgroups on the basis of gender (the number of categories is higher and hence the support is lower). The confidence of {Chemistry, Comprehensive school} ⇒ {Biology} is only slightly lower than {Chemistry} ⇒ {Biology} (0.7000 as opposed to 0.7142) but there is more of a difference in the reverse direction: confidence of {Biology, Comprehensive school} ⇒ {Chemistry} is 0.5295, compared to 0.5727 for all students. For the third rule {Comprehensive school, Male student,

Mathematics} ⇒ {Physics}, we have already seen that the support of the rule {Male student, Mathematics} ⇒ {Physics} is 0.5099 and thus the addition of Comprehensive schools to the LHS makes little difference.

Looking at the rules involving mean GCSE groups, again we obtain some basic insights that Mathematics, Chemistry and Biology are more popular among students with the best GCSE results, but making sense of the longer rules is more complex. The largest difference from the confidence of the equivalent rule without the mean GCSE condition (from Table 2) is found for {Biology, high mean GCSE} ⇒ {Chemistry}, which has confidence 0.6834 as opposed to 0.5727 for all students studying Biology.

Table 15: Top 10 school type ⇒ subject by cosine

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
SUBJECT=Chemistry, schooltype=Comprehensive	SUBJECT=Biology	0.0626	0.7000	2.6165	0.4046
SUBJECT=Biology, schooltype=Comprehensive	SUBJECT=Chemistry	0.0626	0.5295	2.4687	0.3930
GENDER=M,SUBJECT=Mathematics, schooltype=Comprehensive	SUBJECT=Physics	0.0438	0.5217	3.3296	0.3820
SUBJECT=Mathematics, schooltype=Comprehensive	SUBJECT=Physics	0.0550	0.3860	2.4632	0.3682
schooltype=Comprehensive	SUBJECT=General Studies	0.1257	0.2785	1.0779	0.3681
schooltype=Comprehensive	SUBJECT=Psychology	0.1288	0.2855	1.0502	0.3678
schooltype=Comprehensive	SUBJECT=Mathematics	0.1426	0.3161	0.9267	0.3635
SUBJECT=Biology, SUBJECT=Mathematics, schooltype=Comprehensive	SUBJECT=Chemistry	0.0396	0.7113	3.3163	0.3622
SUBJECT=Physics, schooltype=Comprehensive	SUBJECT=Mathematics	0.0550	0.8128	2.3831	0.3622
mean_gcse=High, schooltype=Comprehensive	SUBJECT=Mathematics	0.0723	0.5847	1.7144	0.3520

Table 16: Top 10 mean GCSE ⇒ subject by cosine

LHS	*RHS*	*support*	*confidence*	*lift*	*cosine*
mean_gcse=High	SUBJECT=Mathematics	0.1870	0.5726	1.679	0.5603
SUBJECT=Biology, mean_gcse=High	SUBJECT=Chemistry	0.0950	0.6834	3.186	0.5501
SUBJECT=Chemistry, mean_gcse=High	SUBJECT=Biology	0.0950	0.7289	2.724	0.5087
SUBJECT=Biology, SUBJECT=Mathematics, mean_gcse=High	SUBJECT=Chemistry	0.0684	0.7970	3.716	0.5042
SUBJECT=Mathematics, mean_gcse=High	SUBJECT=Chemistry	0.0995	0.5324	2.482	0.4971
mean_gcse=High	SUBJECT=Chemistry	0.1303	0.3990	1.860	0.4923
SUBJECT=Chemistry, mean_gcse=High	SUBJECT=Mathematics	0.0995	0.7640	2.240	0.4722
mean_gcse=High	SUBJECT=Biology	0.1390	0.4256	1.591	0.4702
GENDER=M,SUBJECT=Mathematics, mean_gcse=High	SUBJECT=Physics	0.0554	0.5717	3.648	0.4497
SUBJECT=Mathematics, mean_gcse=High	SUBJECT=Physics	0.0767	0.4104	2.619	0.4483

4.3 Results: contrasting rules

The previous section has shown that it is hard to derive information on the additional effect of background variables by restricting the form of the rules in this way. In the analysis presented here, we impose more of a structure on what we are looking for: differences in confidence of {subject ⇒ subject} rules across subgroups defined by background variables. The following tables contain comparisons of rules across subgroups defined by three variables: segregated by school type, gender and mean GCSE. The confidence of each rule is reported for each applicable subgroup, and only rules where the confidence varies by more than 0.05 are presented here. In addition, only rules where support (across the whole dataset) is greater than a certain value are reported, in order to keep the results manageable and meaningful.

The rules are presented in Tables 17–19 ordered by the consequent (right hand side of the rule). The confidence values for each subgroup have been shaded according to their difference from the confidence of the rule across the whole dataset: red shading denotes that the confidence is higher; blue shading denotes that the confidence is lower.

Table 17: Contrasts across gender subgroups (support >= 0.02)

LHS	RHS	Total support	F.conf	M.conf
Chemistry	Biology	0.153	0.817	0.622
Chemistry, General Studies	Biology	0.047	0.831	0.643
Chemistry, General Studies, Mathematics	Biology	0.030	0.805	0.582
Chemistry, Mathematics	Biology	0.099	0.788	0.565
Chemistry, Mathematics, Physics	Biology	0.032	0.613	0.425
Chemistry, Physics	Biology	0.042	0.674	0.467
General Studies ,Mathematics	Biology	0.039	0.516	0.351
Mathematics	Biology	0.133	0.475	0.325
Mathematics, Physics	Biology	0.041	0.425	0.282
Physical Education/Sports Studies	Biology	0.029	0.454	0.345
Physics	Biology	0.055	0.483	0.311
Biology, Mathematics, Physics	Chemistry	0.032	0.825	0.752
Biology, Physics	Chemistry	0.042	0.818	0.735
General Studies, Physics	Chemistry	0.023	0.608	0.487
Mathematics, Physics	Chemistry	0.068	0.573	0.500
Physics	Chemistry	0.080	0.586	0.488
General Studies	English Literature	0.052	0.277	0.113
History	English Literature	0.068	0.439	0.236
Mathematics	English Literature	0.033	0.147	0.058
Media/Film/TV Studies	English Literature	0.021	0.251	0.123
English Language	History	0.021	0.177	0.268
English Literature	History	0.068	0.328	0.418
Religious Studies	History	0.025	0.248	0.324
Sociology	History	0.023	0.149	0.223
Critical Thinking	Mathematics	0.029	0.418	0.575
French	Mathematics	0.021	0.331	0.455
General Studies	Mathematics	0.094	0.283	0.454
Psychology	Physical Education/Sports Studies	0.024	0.057	0.162
Biology	Physics	0.055	0.122	0.312
Biology, Chemistry	Physics	0.042	0.179	0.387
Biology, Chemistry, Mathematics	Physics	0.032	0.205	0.450
Biology, Mathematics	Physics	0.041	0.188	0.442
Chemistry	Physics	0.080	0.217	0.515
Chemistry, General Studies	Physics	0.023	0.209	0.502
Chemistry, Mathematics	Physics	0.068	0.264	0.599
Chemistry, Mathematics (Further)	Physics	0.021	0.493	0.784
Chemistry, Mathematics, Mathematics (Further)	Physics	0.021	0.493	0.784
General Studies	Physics	0.045	0.075	0.288
General Studies, Mathematics	Physics	0.037	0.204	0.532
Mathematics	Physics	0.131	0.210	0.510
Mathematics (Further)	Physics	0.038	0.459	0.738
Mathematics, Mathematics (Further)	Physics	0.038	0.459	0.738
Biology	Psychology	0.081	0.386	0.202

LHS	RHS	Total support	F.conf	M.conf
Biology, Chemistry	Psychology	0.034	0.286	0.145
Biology, General Studies	Psychology	0.025	0.396	0.206
Biology, Mathematics	Psychology	0.023	0.245	0.105
Business Studies: Single	Psychology	0.031	0.353	0.186
Chemistry	Psychology	0.041	0.270	0.117
English Language	Psychology	0.034	0.389	0.234
General Studies	Psychology	0.077	0.387	0.196
Geography	Psychology	0.025	0.273	0.131
History	Psychology	0.047	0.293	0.169
Law	Psychology	0.025	0.480	0.297
Mathematics	Psychology	0.055	0.244	0.100
Physical Education/Sports Studies	Psychology	0.024	0.428	0.269

Table 18: Contrasts across mean GCSE subgroups (support>=0.03)

LHS	RHS	Total support	Low.conf	Med.conf	High.conf
English Literature	Biology	0.035	0.068	0.170	0.250
General Studies	Biology	0.080	0.114	0.293	0.451
General Studies, Mathematics	Biology	0.039	0.214	0.353	0.480
Geography	Biology	0.039	0.144	0.294	0.409
History	Biology	0.040	0.087	0.190	0.264
Mathematics	Biology	0.133	0.211	0.346	0.459
Psychology	Biology	0.081	0.142	0.321	0.469
Mathematics	Business Studies: Single	0.033	0.193	0.139	0.051
Biology	Chemistry	0.153	0.366	0.466	0.683
Biology, General Studies	Chemistry	0.047	0.331	0.456	0.695
Biology, Mathematics, Physics	Chemistry	0.032	0.607	0.664	0.827
Biology, Physics	Chemistry	0.042	0.587	0.655	0.826
Biology, Psychology	Chemistry	0.034	0.302	0.374	0.505
General Studies	Chemistry	0.064	0.058	0.185	0.422
General Studies, Mathematics	Chemistry	0.043	0.214	0.348	0.551
Mathematics	Chemistry	0.149	0.214	0.351	0.532
Mathematics, Physics	Chemistry	0.068	0.363	0.418	0.580
Physics	Chemistry	0.080	0.351	0.417	0.586
Psychology	Chemistry	0.041	0.056	0.144	0.277
History	English Literature	0.068	0.230	0.324	0.410
Psychology	English Literature	0.056	0.156	0.218	0.258
Business Studies: Single	General Studies	0.033	0.226	0.293	0.353
English Language	General Studies	0.034	0.263	0.363	0.414
Psychology	General Studies	0.077	0.222	0.299	0.347
Sociology	General Studies	0.033	0.209	0.273	0.333
English Literature	History	0.068	0.213	0.323	0.452
Psychology	History	0.047	0.122	0.179	0.225
Biology	Mathematics	0.133	0.248	0.380	0.618
Biology, Chemistry	Mathematics	0.099	0.401	0.529	0.720

LHS	RHS	Total support	Low.conf	Med.conf	High.conf
Biology, General Studies	Mathematics	0.039	0.215	0.366	0.605
Business Studies: Single	Mathematics	0.033	0.153	0.281	0.459
Chemistry	Mathematics	0.149	0.457	0.593	0.764
Chemistry, General Studies	Mathematics	0.043	0.422	0.572	0.745
Economics	Mathematics	0.057	0.322	0.494	0.706
English Literature	Mathematics	0.033	0.044	0.111	0.283
General Studies	Mathematics	0.094	0.115	0.304	0.570
Geography	Mathematics	0.040	0.110	0.256	0.469
History	Mathematics	0.048	0.065	0.175	0.372
Psychology	Mathematics	0.055	0.085	0.192	0.354
Mathematics	Mathematics (Further)	0.059	0.054	0.089	0.225
Mathematics, Physics	Mathematics (Further)	0.038	0.109	0.155	0.361
Physics	Mathematics (Further)	0.038	0.071	0.121	0.318
General Studies	Physics	0.045	0.052	0.151	0.270
Biology	Psychology	0.081	0.385	0.396	0.232
Biology, Chemistry	Psychology	0.034	0.318	0.318	0.172
Chemistry	Psychology	0.041	0.275	0.275	0.146
English Literature	Psychology	0.056	0.317	0.359	0.222
General Studies	Psychology	0.077	0.306	0.362	0.237
History	Psychology	0.047	0.268	0.296	0.176
Mathematics	Psychology	0.055	0.197	0.217	0.130
General Studies	Sociology	0.033	0.226	0.144	0.052
Psychology	Sociology	0.057	0.309	0.204	0.093

Table 19: Contrasts across school type subgroups (support>=0.04)

			Confidence by subgroup							
LHS	RHS	Total. supp	Acad.	Comp.	FE Coll.	Gram.	Ind.	Sec Mod.	6thF Coll.	Tert Coll.
Chemistry, General Studies	Biology	0.047	0.746	0.697	0.859	0.749	0.755	0.723	0.764	0.733
Chemistry, Physics	Biology	0.042	0.569	0.518	0.584	0.570	0.495	0.565	0.517	0.534
General Studies	Biology	0.080	0.291	0.280	0.272	0.428	0.400	0.247	0.291	0.275
Mathematics	Biology	0.133	0.370	0.390	0.364	0.462	0.368	0.327	0.375	0.371
Mathematics, Physics	Biology	0.041	0.357	0.320	0.364	0.359	0.295	0.371	0.286	0.300
Physics	Biology	0.055	0.393	0.357	0.387	0.402	0.336	0.407	0.313	0.335
Psychology	Biology	0.081	0.264	0.292	0.247	0.405	0.342	0.217	0.300	0.282
Biology	Chemistry	0.153	0.584	0.530	0.551	0.659	0.632	0.435	0.595	0.538
Biology, General Studies	Chemistry	0.047	0.650	0.517	0.576	0.674	0.673	0.434	0.599	0.486
Biology, Physics	Chemistry	0.042	0.833	0.732	0.757	0.790	0.802	0.677	0.777	0.769
General Studies	Chemistry	0.064	0.253	0.208	0.182	0.385	0.357	0.148	0.228	0.182
General Studies, Mathematics	Chemistry	0.043	0.480	0.427	0.419	0.546	0.541	0.414	0.443	0.463
Mathematics	Chemistry	0.149	0.412	0.428	0.371	0.524	0.467	0.328	0.415	0.395
Mathematics, Physics	Chemistry	0.068	0.592	0.514	0.503	0.559	0.544	0.515	0.472	0.479

			Confidence by subgroup							
LHS	RHS	Total. supp	Acad.	Comp.	FE Coll.	Gram.	Ind.	Sec Mod.	6thF Coll.	Tert Coll.
Physics	Chemistry	0.080	0.576	0.503	0.502	0.557	0.545	0.487	0.471	0.482
Psychology	Chemistry	0.041	0.151	0.137	0.127	0.223	0.183	0.080	0.159	0.135
Mathematics	Economics	0.057	0.115	0.122	0.130	0.198	0.302	0.051	0.163	0.125
History	English Literature	0.068	0.332	0.356	0.260	0.417	0.383	0.316	0.279	0.251
Psychology	English Literature	0.056	0.228	0.230	0.163	0.268	0.242	0.221	0.155	0.155
Biology	General Studies	0.080	0.216	0.298	0.091	0.523	0.132	0.217	0.392	0.059
Biology, Chemistry	General Studies	0.047	0.241	0.291	0.096	0.535	0.140	0.217	0.394	0.053
Chemistry	General Studies	0.064	0.229	0.292	0.084	0.526	0.125	0.216	0.388	0.054
Chemistry, Mathematics	General Studies	0.043	0.243	0.281	0.082	0.500	0.116	0.191	0.397	0.061
English Literature	General Studies	0.052	0.161	0.287	0.112	0.478	0.098	0.170	0.342	0.047
History	General Studies	0.058	0.202	0.314	0.089	0.507	0.106	0.220	0.366	0.065
Mathematics	General Studies	0.094	0.209	0.282	0.072	0.480	0.100	0.151	0.372	0.052
Physics	General Studies	0.045	0.214	0.292	0.068	0.498	0.106	0.232	0.381	0.059
Psychology	General Studies	0.077	0.174	0.297	0.083	0.524	0.151	0.176	0.340	0.052
English Literature	History	0.068	0.290	0.351	0.267	0.415	0.419	0.259	0.318	0.287
Biology	Mathematics	0.133	0.476	0.471	0.443	0.587	0.558	0.372	0.485	0.425
Biology, Chemistry	Mathematics	0.099	0.607	0.632	0.578	0.703	0.698	0.576	0.632	0.574
Chemistry	Mathematics	0.149	0.647	0.683	0.620	0.744	0.755	0.617	0.679	0.630
Economics	Mathematics	0.057	0.566	0.548	0.476	0.680	0.642	0.576	0.565	0.519
General Studies	Mathematics	0.094	0.361	0.320	0.262	0.498	0.461	0.196	0.358	0.278
Geography	Mathematics	0.040	0.268	0.306	0.257	0.454	0.343	0.215	0.305	0.269
History	Mathematics	0.048	0.203	0.236	0.148	0.326	0.294	0.146	0.218	0.158
Psychology	Mathematics	0.055	0.165	0.193	0.167	0.299	0.251	0.146	0.201	0.171
Mathematics	Mathematics (Further)	0.059	0.123	0.152	0.132	0.209	0.248	0.121	0.160	0.145
Chemistry	Physics	0.080	0.380	0.381	0.314	0.393	0.424	0.350	0.321	0.338
General Studies	Physics	0.045	0.156	0.158	0.093	0.257	0.234	0.115	0.153	0.138
Biology	Psychology	0.081	0.312	0.318	0.423	0.250	0.136	0.336	0.382	0.404
Chemistry	Psychology	0.041	0.218	0.198	0.299	0.154	0.078	0.203	0.255	0.269
English Literature	Psychology	0.056	0.315	0.318	0.387	0.266	0.133	0.310	0.322	0.352
General Studies	Psychology	0.077	0.277	0.305	0.422	0.264	0.183	0.309	0.321	0.345
History	Psychology	0.047	0.276	0.254	0.335	0.210	0.096	0.287	0.274	0.290
Mathematics	Psychology	0.055	0.152	0.174	0.236	0.145	0.066	0.198	0.197	0.214
Sociology	Psychology	0.057	0.404	0.402	0.500	0.413	0.227	0.389	0.428	0.504
Psychology	Sociology	0.057	0.230	0.197	0.389	0.127	0.023	0.196	0.240	0.274

In Table 17, looking at differences between genders, there is a striking pattern that the difference in confidence is in the same direction for all rules with the same subject on the right hand side. For example, in all rules featuring Biology as a consequent, the confidence of the rule is higher among females than males. This accords with general patterns of uptake of Biology at AS and/or A level (Sutch, 2014, Table 4), where 27.6% of female students took Biology in 2011/12, as

opposed to 25.7% of male students. However, the differences shown here are rather greater: For example, for the rule Chemistry ⇒ Biology, which had the highest support, we find that 81.7% of female students taking Chemistry also studied Biology, compared with 62.2% of male students.

The direction of the difference accords with the uptake statistics in all subjects shown here, with the exception of Chemistry: the confidence of the rules is consistently higher for females than males, but this contrasts with the general pattern in uptake, where the uptake among males was higher (24.5% compared to 18.9%).

Looking at mean GCSE (Table 18), again there is a strong pattern that the difference in confidence is in the same direction for all rules with the same subject as consequent. For example, students from the higher attainment group were more likely to study Mathematics, given they were also studying the subjects on the left hand side, than students from the lower attainment group. This accords with the general patterns of uptake (Sutch, 2014, Table 5). For most rules the group with the highest confidence is the highest attainment group. This is to be expected as these students take a larger number of subjects on average.

The pattern is less clear for school type (Table 19), although there are some subjects where the rules are more consistent (General Studies having the highest uptake in grammar schools, psychology having highest uptake in FE Colleges). For example, the rule {Psychology ⇒ Biology} had much higher confidence in grammar schools (0.405) than academies (0.264), reflecting the uptake of A level[7] Biology at these centre types, but the rule Physics ⇒ Biology has a similar confidence for each (0.402 and 0.393 respectively). It is not immediately clear why this should be, but this seems to point to differences in the type of students studying Psychology at Academies or Grammar schools.

5 Discussion

This research has shown that association rules can be used to investigate patterns of subject uptake by students, and that deep relationships between subjects are uncovered.

Association rules provide a framework for inspecting relationships, not only the most obvious but also some beneath the surface, through a choice of different interestingness measures. Because the topic of student subject choice has been investigated before, not least in the annual Statistical Reports, we have already accumulated a body of knowledge which serves as a hindrance as well as a help: many of the patterns that were uncovered were already known to us, and there were few truly novel results. Thus most of the generated rules would have scored less well against more semantic measures of interest proposed by Geng and Hamilton (2006). Cutting through the ‘known’ or ‘obvious’ results was difficult and time-consuming. An interactive method of filtering the rules may have proved fruitful.

One disadvantage of association rules is that taking the results further and probing deeper requires drawing comparisons between rules (whether implicitly or explicitly). The measure of lift provides a way to do this against the null situation, where the LHS of the rule is empty, but this does not allow gauging the additional effect when particular conditions are added to or removed from the LHS of the rule. Contrast sets/subgroup discovery may be promising but would require further work to investigate, and it is likely that this effort would be better spent on other modelling techniques. Perhaps the greatest value of association rules is in generating hypotheses for other research, by presenting ‘interesting’ rules for the researcher to inspect, and then explore in more depth in separate investigations.

The strongest associations between subjects, reflecting all previous analysis of AS/A level subject combinations, are between Science subjects (including Mathematics and Further

[7] Sutch (2014) does not present analyses by centre type for AS levels.

Mathematics). A student is much more likely to be taking a Science subject at AS if he/she is already studying another one. These Science rules had strong support, lift and cosine.

However, there were many rules with high lift but less strong support and confidence, of which the most prominent were associations between classical subjects. Students studying Latin were much more likely than average to be taking Classical Greek (but still unlikely on an absolute level).

Cosine did not prove to be a useful interestingness measure for student subject choice, as it places too much emphasis on support: an item with very high support (for example, the choice of Mathematics A level, or the fact of a student being male) seems to dominate the list, even if the confidence of the rule is low, and the lift very close to 1. The support of items in our dataset tended to be small, and the advice of Merceron & Yacef (2008) to use cosine may therefore not apply: as Merceron and Yacef (2008, pp. 4–5) say, "added value and lift rely on probabilities, which make more sense when the number of observations is large" and as such lift may be more appropriate here.

Data on combinations of subjects has been presented before; for example, Gill (2012) showed figures for uptake of pairs of subjects by the whole cohort, and then broken down by subgroup. However, comparison of these figures becomes unwieldy due to the number of dimensions involved. Section 4.3 shows how association rules could be constrained and compared, but this is still not wholly satisfactory. The results obtained by considering subgroups are not too surprising as they mostly follow from the results at a single subject level, suggesting that the addition of subject combinations is not adding a great deal. However, some of the differences highlighted across subgroups are more dramatic than at a single subject level, for example the split between gender in Science subjects. Moreover, the gender difference in Chemistry uptake actually reverses if one measures uptake of Chemistry among students who are taking Physics.

The differences by subgroup do serve as a reminder that such differences exist, even if they are not attributable to the combinations. These differences could have implications for the uptake of certain subjects at university if pre-requisites for a certain course are taken to a greater or lesser extent. For example, if female students taking Psychology A level are less likely to be studying Mathematics at the same time, this may bar their entrance from certain university courses. Indeed, there may even be implications for assessment and awarding at A level if certain subgroups are disadvantaged by not taking a related subject.

The constrained nature of students' choices at AS/A level may have some implications for the applicability of association rules – in particular the reality that most students can only study four subjects at most due to timetabling constraints. It may be useful to consider a modelling approach that recognises this fact, such as discrete choice models.

In this research, data on prior attainment at GCSE has only been included in a grouping at the student level based on mean GCSE. Grades in each GCSE subject have not been used, but this certainly has a strong link to uptake of individual subjects at AS and A level (Sutch, 2013), so further work could look at the detailed influence of prior attainment on uptake of certain combinations.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD record* (Vol. 22, pp. 207–216). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=170072

Bay, S. D., & Pazzani, M. J. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, *5*(3), 213–246. doi:10.1023/A:1011429418057

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, *14*(2), 77–88. doi:10.1016/j.iheduc.2010.07.006

García, E., Romero, C., Ventura, S., de Castro, C., & Calders, T. (2009). Association rule mining in learning management systems. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d Baker (Eds.), *Handbook of educational data mining*. Boca Raton, FL: Chapman & Hall/CRC.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, *38*(3). doi:10.1145/1132960.1132963

Gill, T. (2012). *Uptake of two-subject combinations of the most popular A levels in 2010, by candidate and school characteristics* (Statistics Report Series No. 38). Cambridge, UK: Cambridge Assessment.

Gill, T. (2014a). *Provision of GCE A level subjects 2013* (Statistics Report Series No. 73). Cambridge, UK: Cambridge Assessment.

Gill, T. (2014b). *Uptake of GCE A level subjects 2013* (Statistics Report Series No. 72). Cambridge, UK: Cambridge Assessment.

Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2014). *Arules: Mining association rules and frequent itemsets*. Retrieved from http://CRAN.R-project.org/package=arules

Kralj Novak, P., Lavřac, N., & Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, *10*, 377–403. Retrieved from http://jmlr.org/papers/v10/kralj-novak09a.html

Kumar, V., & Chadha, A. (2012). Mining association rules in student's assessment data. *International Journal of Computer Science Issues*, *9*(5), 211–216 .

Luna Bazaldua, D. A., Baker, R. S., & San Pedro, M. O. Z. (2014). Comparing expert and metric-based assessments of association rule interestingness. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 44–51). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/44_EDM-2014-Full.pdf

Merceron, A., & Yacef, K. (2008). Interestingness measures for associations rules in educational data. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 57–66). Retrieved from http://www.educationaldatamining.org/EDM2008/uploads/proc/6_Yacef_18.pdf

Minaei-Bidgoli, B., Tan, P.-N., & Punch, W. F. (2004). Mining interesting contrast rules for a web-based educational system. In *2004 International Conference on Machine Learning and Applications, 2004: Proceedings* (pp. 320–327). doi:10.1109/ICMLA.2004.1383530

Pechenizkiy, M., Trčka, N., De Bra, P., & Toledo, P. A. (2012). CurriM: Curriculum mining. In Yacef, K., Zaïane, O., Hershkovitz, H, Yudelson, M., & Stamper, J (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 216–217). Chania, Greece. Retrieved from http://educationaldatamining.org/EDM2012/uploads/procs/Posters/edm2012_poster_11.pdf

Power, D. J. (2002). Ask Dan! - What is the “true story” about data mining, beer and diapers? *DSS News*, *3*(23). Retrieved from http://www.dssresources.com/newsletters/66.php

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(6), 601–618. doi:10.1109/TSMCC.2010.2053532

Romero, C., Romero, J. R., Luna, J. M., & Ventura, S. (2010). Mining rare association rules from e-learning data. In *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 171–180). Retrieved from http://sci2s.ugr.es/keel/pdf/keel/congreso/RareAssociationRuleMining.pdf

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. d. (2011). *Handbook of educational data mining*. Boca Raton, FL: Chapman & Hall/CRC.

Russell Group. (2011). *Informed Choices: A Russell Group guide to making decisions about post-16 education*. London: Russell Group. Retrieved from http://www.russellgroup.ac.uk/informed-choices.aspx

Singleton, A. D. (2009). Data mining course choice sets and behaviours for target marketing of higher education. *Journal of Targeting, Measurement and Analysis for Marketing*, *17*(3), 157–170. doi:10.1057/jt.2009.13

Sutch, T. (2013). *Progression from GCSE to AS and A level, 2010* (Statistics Report Series No. 69). Cambridge, UK: Cambridge Assessment.

Sutch, T. (2014). *Uptake of GCE AS level subjects 2007–2013* (Statistics Report Series No. 75). Cambridge: Cambridge Assessment.

Tan, P.-N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293–313. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306437903000723

Vialardi, C., Bravo, J., Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. *International Working Group on Educational Data Mining*. Retrieved from http://eric.ed.gov/?id=ED539088