

Using SAS Enterprise Miner to Examine General Education Issues

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

General education courses are required to give students a broad background in a variety of disciplines. However, outside of freshmen English and possibly history, many different courses are available to satisfy these requirements. It is rare to examine how students navigate through all of the available course offerings to complete general education. All student enrollment from 2000-2004 was made available for data mining along with student demographics, including ACT scores. With 22,000 students enrolled in any given semester averaging 15 credit hours, the database is extensive. Sequential market basket analysis was used to examine pathways taken by students through general education, and to determine which pathways are more successful compared to others. SAS Text Miner is also used to investigate general education pathways. Enterprise Miner is combined with SAS/Stat to drill down into the data. Results show that course choice does impact student success.

INTRODUCTION

The number of different combinations of courses that are possible for undergraduate majors is almost unlimited. In order to examine student pathways to success, it is necessary to examine all combinations. Then the problem is very similar to that of examining customer purchases when there are thousands of possible purchase items and purchase combinations. However, unlike customer purchases, there is a final outcome variable for students with success measured by grades and by graduation. When there are thousands of possible combinations available for student choice, transaction counts become so small that the results often become highly questionable, even though choice of major does tend to restrict some combinations. Attempts have been made to address the issue through the development of clusters that can be used to reduce the number of possible course combinations to be examined. However, any time data compression occurs, information is lost. Another means of reducing the problem is to reduce the data to a subset of items. This will allow the intense examination of items within subsets, but will not provide information on how items are related across subsets.

Market basket analysis can be combined with other data mining and statistical tools to drill down into student preferences to examine pathways of choice. Sequential path analysis will be used since students enroll in courses sequentially by semesters. It is the purpose of this paper to discuss an application where student preference can lead to success or failure in outcomes. The first step is to examine the nature of student preferences given choice of major. The initial focus will be on general education courses that are required of all students at the University of Louisville. These include English, History, and Communications. Students are also required to take a mathematics course, although many students must first enroll in remedial mathematics before advancing to the level of general education.

Another technique used for compression of course combinations will be text analysis. Text strings that contain all courses by individual student will be created, and those text strings will be clustered using SAS Text Miner. The SAS code needed to create the text strings is given in the appendix. It will be shown that students taking specific mathematics general education courses will have a greater likelihood of graduating compared to students taking different courses. Certain pathways are more likely to lead to success.

METHOD

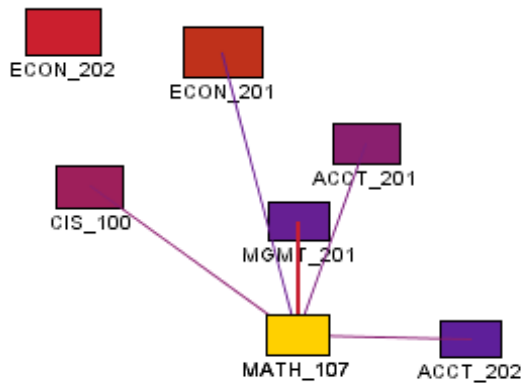
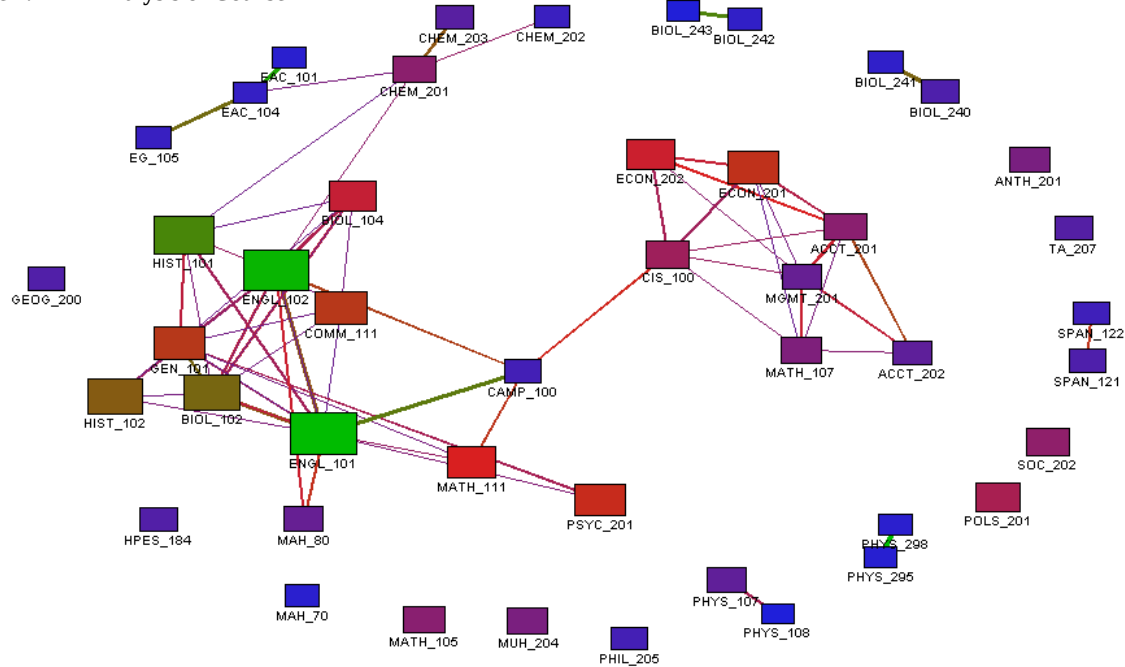
The dataset studied was the admissions file for all students for the years 2000-2004. All course enrollment information was included in the dataset along with certain admissions information, including ACT and SAT score information. The observational unit was the individual student course enrollment. Therefore, each student is listed in the database for each semester and each enrollment during the time of study. For 22,000 students enrolled in a given year, the database contained information on 54,000+ students and 780,000+ enrollments.

The department was recorded in one field with the course number in a second field. The two fields were concatenated with an underscore between the department name and course number. In that way, the course identity can be maintained while using sequential path analysis in Enterprise Miner, version 5. The SAS statement used to combine these two fields, new for Version 9, is equal to `CATX('_', course name, course number);`. The first entry gives the character used to separate the concatenated items that are listed.

RESULTS

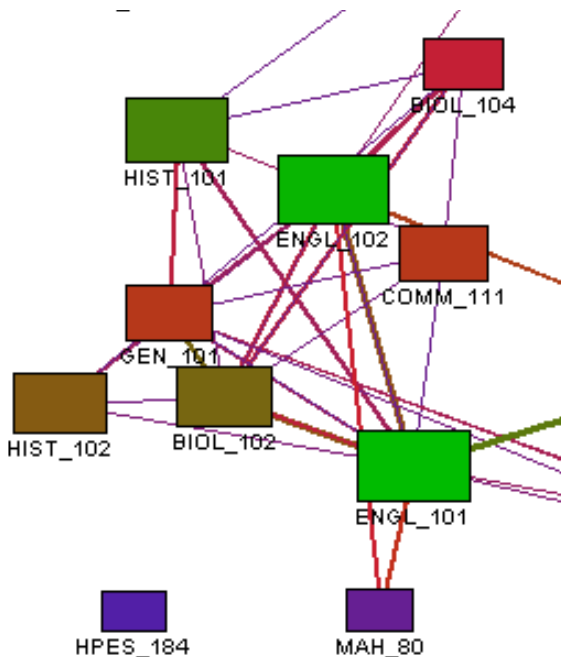
A sequencing variable was added to the enrollment dataset, limited to general education courses defined with numbers 100-200. The sequencing was used to investigate the pathways students take to complete their general education requirements, and the ordering with which they are taken. A sequential link analysis is given in Figure 1.

Figure 1. Link Analysis of Course



According to the links, students in Business enroll in Economics 201 and 202, Accounting 201 and 202, Management 201, Basic computing (CIS 100, Microsoft Office) and Finite Mathematics (Math 107). There is one link from the Business Cluster to the second cluster and it runs through Camp 100, or Campus Culture for Business Students, required of all Business majors. Camp 100 is part of a triangle that also includes Math 111 (College Algebra) and Psyc 201 (Introduction to Psychology). Math 111 is required of all Business majors unless they substitute a higher-level mathematics course; Psychology fulfills a social science general education requirement.

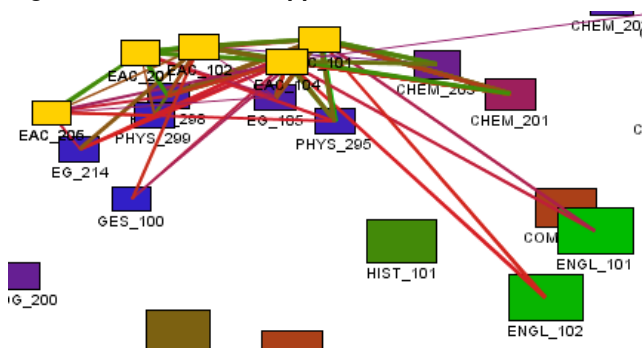
Figure 3 displays the second cluster of courses. Crossover indicates students in Liberal Arts also take these courses (Math 111, Psyc 201, Camp 111).



One of the courses in this cluster is GEN 101, or Arts & Sciences Orientation. It is clear that this cluster is generally for A&S students. English 101 and 102 are required of all students at the University, as is History 101 and 102 and Communications 111. Biology 102 seems to be the science course of choice for A&S majors. Of interest is the course, MAH 80, which is pre-algebra and is required placement for all students who need remedial mathematics to complete their general education. Domain knowledge indicates that A&S and Business have more declared majors than any other professional discipline. For this reason, market basket analysis can only find connections in those majors; information about Engineering students and other schools is not available without some additional pre-processing.

In the upper left hand corner of Figure 1, there are a few EAC courses. The EAC value indicates engineering core courses. However, in comparison to the number of A&S and Business students, there are few engineering students so that their pathways are not represented. More paths can be identified by reducing the support level to increase the number of links (Figure 4).

Figure 4. Decrease in Support



A decrease in support increases the number of links in the link analysis. It shows that most of the connections are to physics and to chemistry as well as connections to the basic general education courses of English, History, and Communication.

Figure 5 shows the connections restricted to Engineering majors. In order to investigate pathways of engineering students, we isolate them from the rest of the data and perform market basket analysis on the restricted dataset. The problem with doing this is that the database does not contain any identifiers as to which students are engineering majors. Therefore, it requires some programming to isolate these majors. The first step is to create text strings of courses, changing the observational unit from course to student. The programming steps (in SAS) are given in the appendix.

The next step is to use the text strings of courses to isolate engineering students by filtering only those who were enrolled in at least one of the core engineering courses. Once the dataset has been reduced for engineering students, the pathways can be examined (Figure 5). While the courses in Figure 5 are very similar to those in Figure 4, additional emphasis is given to the Physics courses as linked to Engineering courses, shown in Figure 5.

The graph displays a complex network of connections between various course combinations. The nodes are labeled with course codes, often followed by an ampersand and another course code, indicating combinations of courses. The nodes are color-coded: yellow, blue, red, and green. The connections are represented by lines of different colors (red, green, blue, yellow) and thicknesses, suggesting different types or strengths of relationships between the course combinations. The graph is highly interconnected, with many nodes having multiple connections to other nodes.

A similar examination of general education mathematics courses was examined using similar programming. Figure 6 shows the pathways from Math 105, Contemporary Mathematics. Note that most lead to courses labeled ELFH, Department of Leadership, Foundations and Human Resource Education for education administrators. The other courses listed are also related to education with the exception of TA, or Theater Arts. For College Algebra (Figure 7), the relationships are to courses in the College of Business, with the exception of one connection to Psychology and one to English.

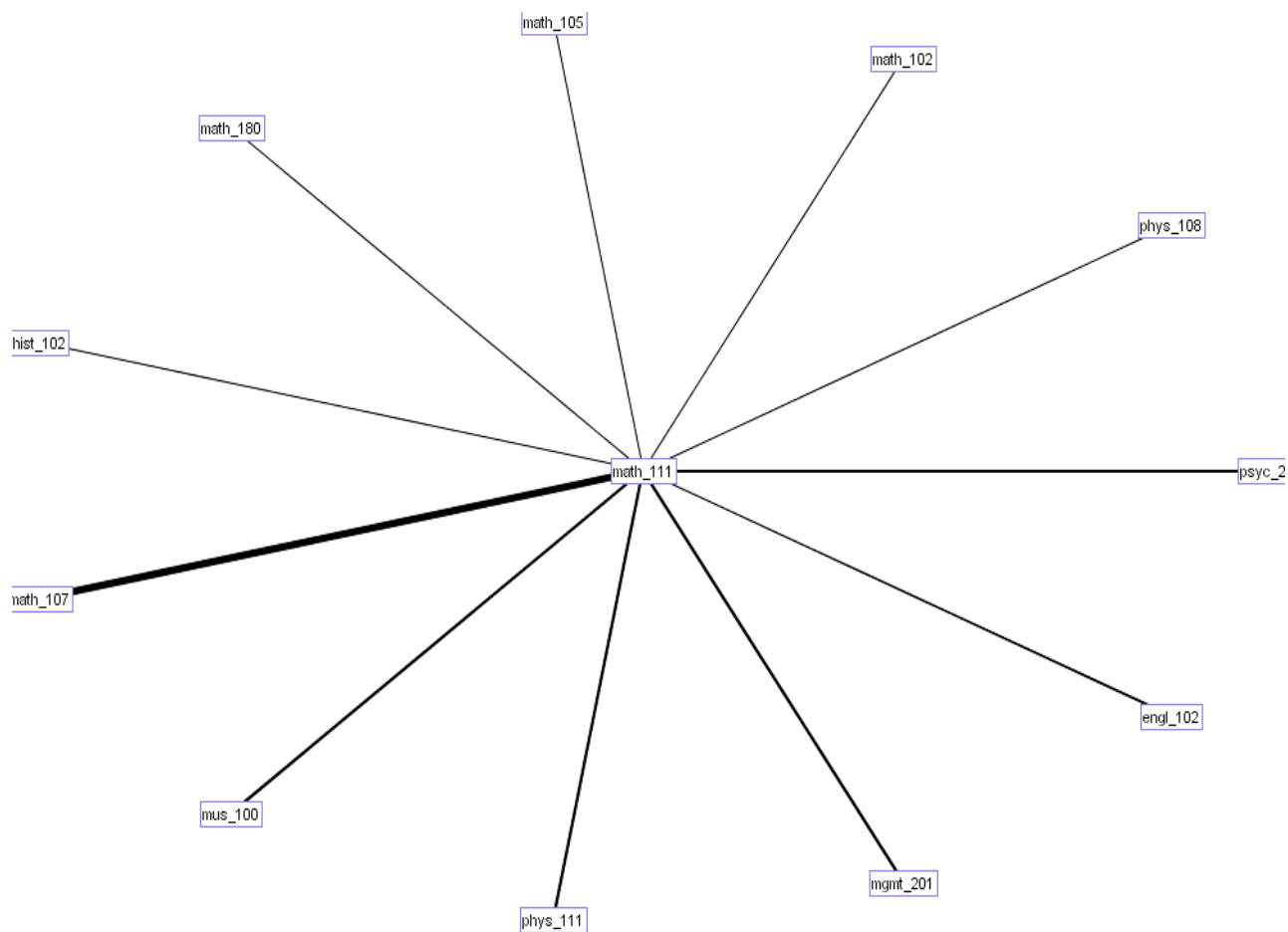
4

[illegible]

TEXT ANALYSIS

College Algebra

Figure 8. Results of Text Analysis



RELATIONSHIP TO OUTCOMES

Once the pathways have been discovered, the relationship between pathways and outcomes can be examined. It is clear from the link analyses that students in different academic disciplines choose different general education mathematics courses. As discussed previously, students enrolled in Math 105, Contemporary Mathematics are strongly associated with education courses; those enrolled in Math 111, College Algebra, are strongly associated with Business courses. It should be noted that Math 105 is not an algebra-based course.

There are three fields that should be considered in relationship to the student's chosen path through general education. The first is the score on the entrance examinations in relationship to course. The second field contains all grades assigned in the course. The final field is graduation. In many cases, students take more than one mathematics course, especially if they must start at a remedial level. In order to capture all of the mathematics pathways, text strings restricted to the mathematics courses were constructed, and used to define clusters of students. Once the text strings were defined (see appendix) and clustered using expectation maximization, there were a total of 10 clusters as labeled in Table 1.

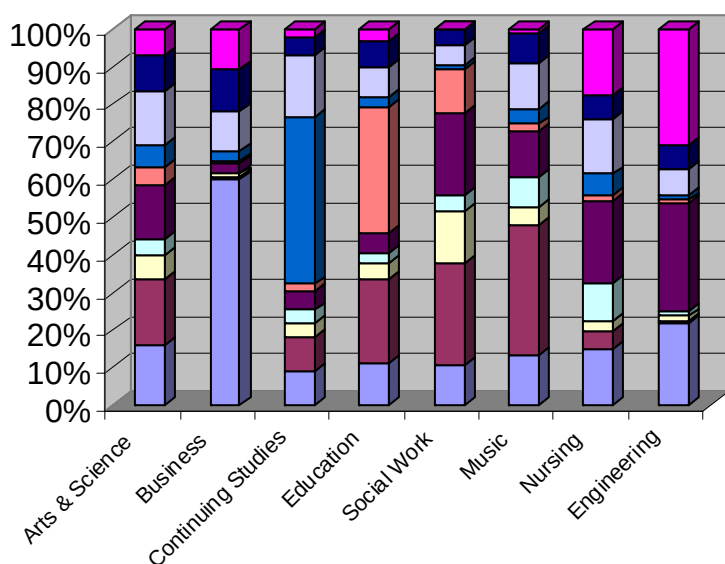
Table 1. Clusters of Mathematics Courses for Enrolled Students

Cluster Number	Mathematics Courses
1	Trigonometry, Precalculus, Elementary Statistics, College Algebra, Elementary Calculus
2	Remedial Mathematics, Finite Mathematics, College Algebra, Intermediate Algebra
3	Remedial mathematics, Contemporary Mathematics, College Algebra
4	Remedial mathematics, Contemporary Mathematics

Cluster Number	Mathematics Courses
5	Mathematics for Teachers, Elementary Statistics, Remedial Mathematics
6	Elementary Statistics, Elementary Calculus, Precalculus, Mathematics for Teachers
7	Elementary Statistics, Contemporary Mathematics, Intermediate Algebra
8	Contemporary Mathematics, Remedial mathematics, Intermediate Algebra, College Algebra
9	Contemporary Mathematics
10	College Algebra, Finite Mathematics

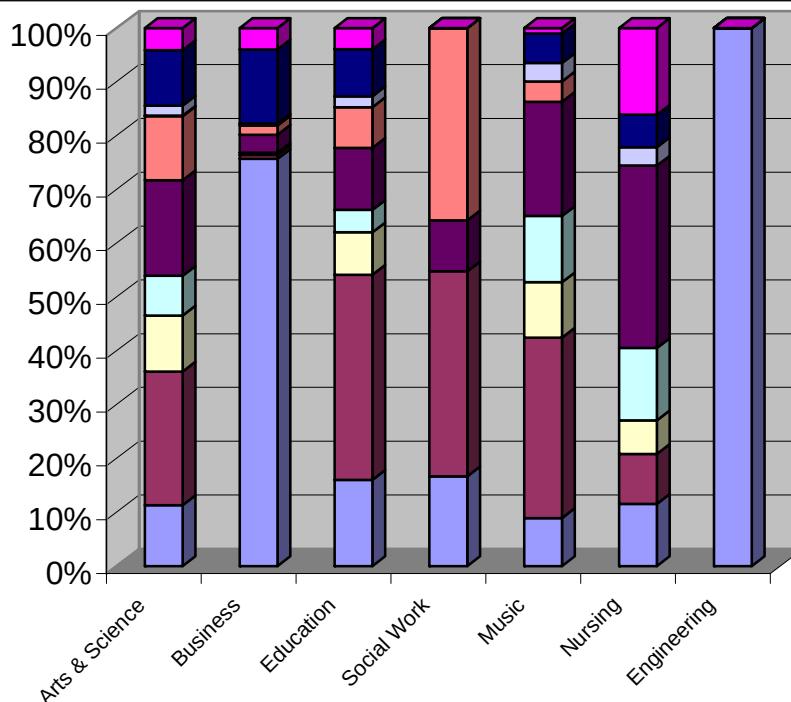
Note that cluster 10 corresponds to the most common pathway into the College of Business; cluster 9 corresponds to Education. In this text analysis, only 100 level mathematics courses were included; students beginning in Calculus were not part of this analysis. The proportion of students in each of the clusters by school is given in Figure 9; the likelihood of graduation is given in Figure 10.

Figure 9. Clusters of Mathematics Courses by School



- Trigonometry, Precalculus Elementary Statistics, College Algebra, Elementary Calculus
- Remedial, Finite Mathematics, College Algebra, Intermediate Algebra
- Remedial, Contemporary Mathematics, College Algebra
- Remedial, Contemporary Mathematics
- Mathematics for Teachers, Elementary Statistics, Remedial
- Elementary Statistics, Elementary Calculus, Precalculus, Mathematics for Teachers, Remedial
- Elementary Statistics, Contemporary Mathematics, Intermediate Algebra, Elementary Calculus, Remedial
- Contemporary Mathematics, Remedial, Intermediate Algebra, College Algebra
- Contemporary Mathematics
- College Algebra, Finite Mathematics

Figure 10. Mathematics Cluster for Graduating Students

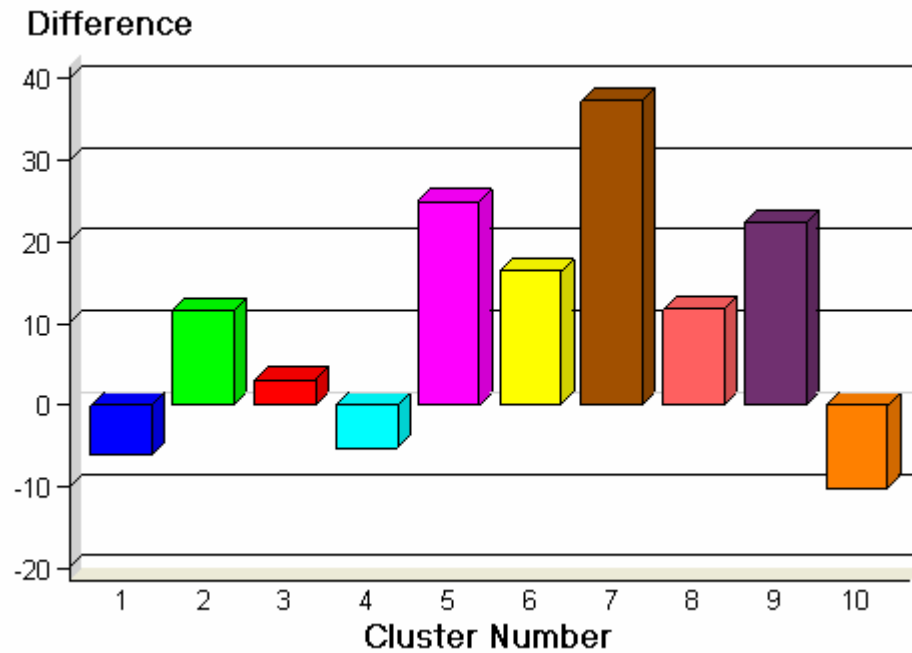


- Trigonometry, Precalculus Elementary Statistics, College Algebra, Elementary Calculus
- Remedial, Finite Mathematics, College Algebra, Intermediate Algebra
- Remedial, Contemporary Mathematics, College Algebra
- Remedial, Contemporary Mathematics
- Mathematics for Teachers, Elementary Statistics, Remedial
- Elementary Statistics, Elementary Calculus, Precalculus, Mathematics for Teachers, Remedial
- Elementary Statistics, Contemporary Mathematics, Intermediate Algebra, Elementary Calculus, Remedial
- Contemporary Mathematics, Remedial, Intermediate Algebra, College Algebra
- Contemporary Mathematics
- College Algebra, Finite Mathematics

While the link analysis shows that the most common path into Business is via College Algebra, 40% start at remedial mathematics. Similarly, more students in Education enroll in Mathematics for Teachers; however those who enroll in the specific discipline of ELFH take the Contemporary Mathematics route. Of interest also is the fact that some students in Engineering also start at the College Algebra level. Social Work, Music, and Nursing were not present in the link analyses; the total enrollment is very small compared to other disciplines so that transaction count is low for their students.

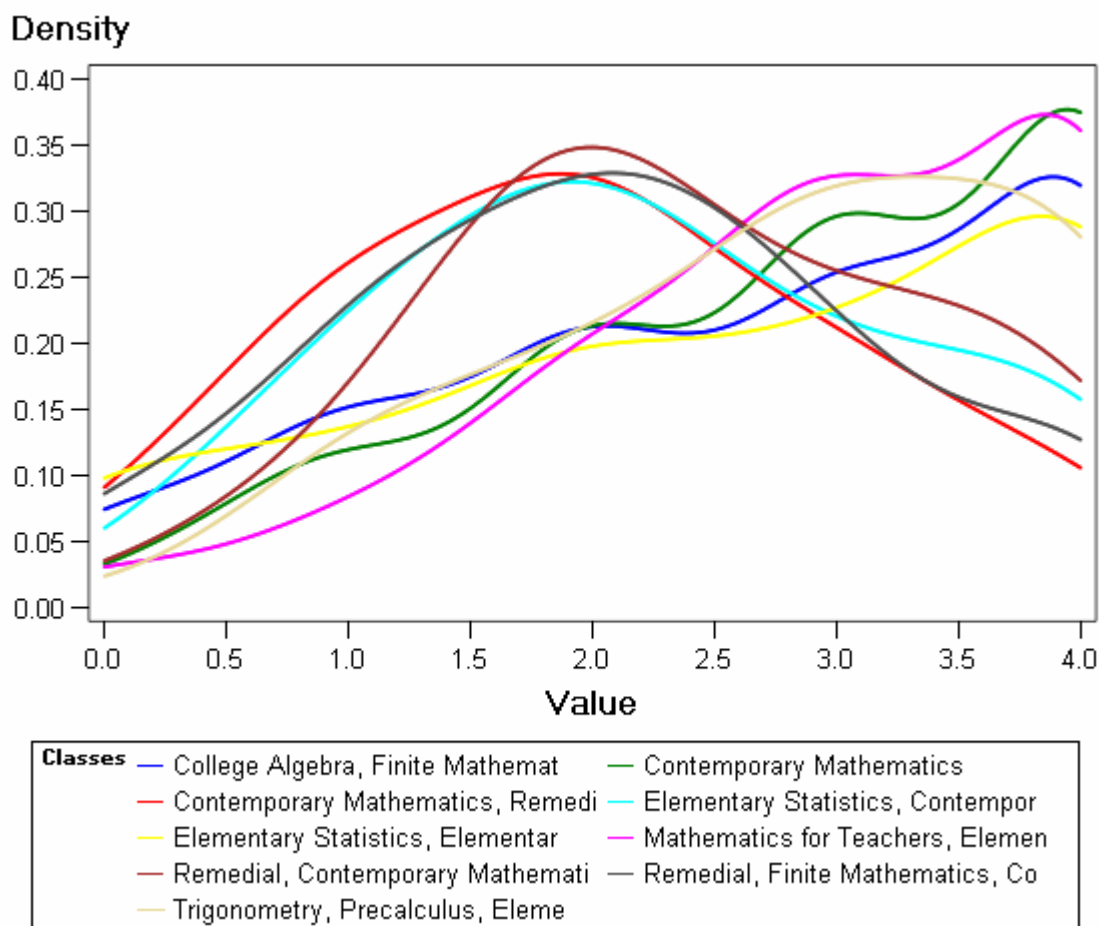
Figure 10 gives the relationship of cluster limited to students who graduate in their given disciplines. It shows that engineering students who start in any of clusters 1-9 do not graduate; the Business students who begin in cluster 10 represent over 70% of the graduates. Students who begin at the remedial level are a much smaller proportion of graduates compared to freshman. The difference between enrollment cluster and graduation cluster is given in Figure 11.

Figure 11. Difference Between Enrollment and Graduation by Cluster



The clusters can be divided into those that contain mostly algebra-based courses and those that are not algebra-based. Unquestionably, there are two different patterns to graduation, depending upon the grouping (Figure 12). The non-algebra based courses have curves that increase from left to right, indicating that the greater the average grade in the mathematics course, the greater the likelihood of graduation. However, for the non-algebra based courses, the peak value occurs at an average grade of 2.0, indicating that lower grades in these courses can lead to successful graduation.

Figure 12. Average Grade in Mathematics Cluster for Graduating Students



DISCUSSION

There are many situations where there are too many different levels in a data field that must be examined. It is important to maintain linkage between items in the data field by identifier, in this example, student identification number. With student courses, it is not particularly useful to examine courses only across departments; different courses within departments are also of interest. Therefore, other methods need to be used to compress data so that meaningful extracts of information can be found.

The data demonstrated that different student majors are associated with different general education mathematics courses, with differing levels of success. Students who have difficulty in mathematics can have difficulty graduating and should declare a major that is associated with their mathematics abilities. In other words, students who do not have the background for algebra-based courses should declare a major that will allow for a non-algebra general education mathematics course to satisfy the requirements. Otherwise, a student will make the decision to change majors or to leave the university only after failure in the mathematics requirements.

CONTACT INFORMATION

Patricia B. Cerrito
 Department of Mathematics
 University of Louisville
 Louisville, KY 40292
 502-852-6826
 502-852-7132 (fax)
 pcerrito@louisville.edu

APPENDIX

The code listed below is used to change the observational unit from course to student:

```
proc sort data = sasuser.data out= work.sort_out;
  by stuid course;
run;
data work.sort_out1;
  set work.sort_out;
  course = translate(left(trim(course)),'_',' ');
run;
```

Sort the data by student identifier, stuid.

```
proc Transpose data=work.sort_out1
  out=work.tran
  prefix=crs_;
  var course ;
  by stuid
run;
```

Shifts all courses for one student into a series of fields; one observation per student remains.

```
data work.concat( keep= stuid course ) ;
  length course $32767 ;
  set work.tran ;

  array chconcat {*} crs_ ;

  course = left( trim( crs_1 )) ;

do i = 2 to dim( chconcat ) ;
  course = left(trim(charges)) || ' ' || left(trim( chconcat[i] )) ;
end ;
run ;
proc sql ;
  select max( length( course )) into :course_LEN from work.concat ;
quit ;
%put course_LEN=&course_LEN ;
data sasuser.crsstrings ;
  length course $ &course_LEN ;
  set work.concat ;
run ;
proc contents data=sasuser.crsstrings ; run ;
```

Concatenates the fields created in PROC TRANSPOSE, and creates one long text string to contain the all courses per student.

Finds longest text string to be used to define final data.set

The code listed below is used to filter out all engineering students by using the 'Contains' statement:

```
PROC SQL;
  CREATE TABLE SASUSER.QURY5744 AS SELECT
GENEDTEXT.course,
  GENEDTEXT.STUID FORMAT=BEST12.,
  GENEDTEXT.coursestring,
  GENEDTEXT.i
  FROM SASUSER.GENEDTEXT AS GENEDTEXT
WHERE GENEDTEXT.coursestring CONTAINS 'EAC;
QUIT;
```

Once the student identifiers have been filtered, all other courses are merged back into the dataset:

```
PROC SQL;
  CREATE TABLE SASUSER.QURY5845 AS SELECT QURY5744.course,
  QURY5744.STUID FORMAT=BEST12.,
```

```

QUERY5744.coursestring,
QUERY5744.i,
classesmodified.DEPT FORMAT=$F4.,
classesmodified.CRSENO FORMAT=BEST12.,
classesmodified.SECTNO FORMAT=$F4.,
classesmodified.CRHRS FORMAT=BEST12.,
classesmodified.GRADE FORMAT=$F2.,
classesmodified.AUDIT FORMAT=$F1.,
classesmodified.STRM FORMAT=$F4.,
classesmodified.STUID FORMAT=BEST12. AS STUID1,
classesmodified.course FORMAT=$F17. AS course1,
classesmodified.GRADEPOINT FORMAT=BEST12.
FROM SASUSER.QUERY5744 AS QUERY5744
  LEFT JOIN EC100020.classesmodified AS classesmodified ON (QUERY5744.STUID = classesmodified.STUID);

```