# MIDFIELD Special Session; A Primer on Novel Methodologies in Longitudinal Analysis of Student Data

George Ricco
RB. Annis School of Engineering
University of Indianapolis
Indianapolis, IN, USA
riccog@uindy.edu

***Abstract*— Innovative Practice.** The past decade has opened up MIDFIELD to the use of novel methodologies. Three particular phenomena – grade variance due to course size and enrolled section, and degree program change – are of particular interest. This work demonstrates three useful methodologies for analyzing these phenomena, how they are used, and preliminary results. The first, Markov chains, can examine links between *success* (graduation) and changing majors. The second, Hierarchical Linear Models (HLMs) can show nested relationships between different courses, including the effect of course size and section on grade variances. The third, mutual information, can uncover relationships between sets of students who switched majors and those who did not, relative to a variable such as GPA. Some novel results include: course size has no effect on grade variance, but section does; students who switch majors graduate at a higher rate; and the effect of switching majors is more *random* for those who leave college than those who graduate.

***Keywords—MIDFIELD, analytics, big data, data science, information theory, informatics, Markov, HLM, MLM, hazards models, cluster, First-Year Engineering.***

## I. Introduction

The purpose of this paper is to describe the methodologies employed in recent efforts to apply more advanced statistical analysis techniques in traversing the MIDFIELD dataset. Some of these efforts are a few years old, and others are either in review or about to be published [1-6]. Most of these methodologies were employed when the data set was in its second stage of growth, with between nine and eleven institutions fully incorporated, and of course now is on its way to an order-of-magnitude expansion [7, 8].

The first impetus behind this push is to demonstrate to the community the usefulness of these methodologies, and how they can be applied to great success. The second reasoning here is to provide a pathway forward for young and aspiring MIDFIELD researchers to employ new methods otherwise ignores or considered too time-consuming or onerous. The third – and this may be the most subversive – is that most MIDFIELD work until this time has been studying in aggregate relationships between variables. Being that summation or aggregation of categorical variables is the primary way educational reports operate at institutional and state levels, this makes sense. The statistical methodologies a researcher employs must match those that his contemporaries respect and that can be understood by their audience!

## II. Background

We will cover a brief outline of three tools used recently within MIDFIELD to analyze student data. The first being a Markov chain model [9], the second being a hierarchical linear model (HLM) analysis [1, 3], and the third being a information theoretical model based on Shannon entropy [4]. There are other more recent analyses that we will cover in our special topic session, including one on cluster analysis and another on an expanded hazards model treatment – the cluster analysis focusing [5]on first-year courses in engineering and the hazards model focusing on sets of students who switch majors multiple times [6].

### *A. Markov Chains*

Markov chains are a commonly utilized modeling technique for incorporating elements of uncertainty into arbitrary dynamic systems. These models map or append some sort of parameter domain, to an arbitrary state space, the values of which the process can take. The parameter domain may or may not be temporal, but often times is time-based in some fashion. These models are suited for studying long-term, probabilistic behavior of convoluted or coupled systems which cannot be fully characterized by traditional descriptive statistics. We have seen the importance of Markov chain understanding increase as its utilization in research continues to expand.[10] We continue this trend by using this framework to model student attrition and its relationship with major switching.

Reviews of modeling theory in the educational space [11] suggest that such a mathematical construct could provide some relevant information about high level problems with regards to educational systems, hence eschewing the high variability present in building probabilistic models on an institutional level. Hence, to build our chain we employ the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) database to address student attrition on a multi-institutional level.

### B. Hierarchical Linear Models

Perhaps the most *in vogue* amongst modern educational statistical methods is the *hierarchical linear model* or *multi-level modeling* technique. For reasons that are obvious ("it's in the name,) this technique is a hierarchical or nested linear model. The power of such a model is that it can provide a glimpse into the how an outcome is primarily affected by variables endemic to the nested structure of the data itself, or if the bins make no significant difference. Finally and most importantly, the primary assumption of any HLM is *that variance of parameters occurs at more than one level.* Throughout the literature, the name of the game within HLM is *variance.* We will explore this in our example.

As a note on background, these models are known by a plethora of names, such as hierarchal linear models [12], multilevel models [13], generalized linear mixed models [14], nested models [15], mixed models or mixed effects models, random coefficient models [16], random effects models [17], random parameter models [18-21], split plot models; covariance components models, and others. It is important for the reader to know that while the sheer number of names is confusing, the fact the names mean different things outside of the world of HLM further compounds this confusion. Ex. A *split plot model* is not a HLM according to our colleagues in the world of nursing statistics.

Historically, this model's efficacy became mainstream in the infamous *High School and Beyond* survey analysis performed by Coleman, Hoffer, and Kilgore [22, 23]. In this study, the researchers wanted to determine the effect of attending a *public* school versus attending a *private* school on mathematics achievement scores (the outcome). They looked at groups of schools (the clusters or nested models) versus the overall group of students without clusters (aka the grand ensemble and grand mean). What they noticed was stunning – while the *herd* effect of poorly-achieving students being surrounded by overachieving students was realized, it appeared that private schools did a better job of improving the math achievement gradient over time than public schools. This is only one example of many of how HLMs have been employed in educational statistics, but it is by far one of the most discussed.

### C. Mutual Information Theory

A more modern technique (perhaps as popular as compressive sensing within electrical engineering) in data science is *mutual information*. Put simply, this exploits the Shannon entropy of a state or system of states to allow the researcher to make a base conclusion about the *order* or *disorder* of a system. So what makes this approach different than, a chaotic (indeterministic) or fractal (deterministic) approach employing a Kolmogorov factor? First, the results are to determine the overall entropy contained between two sets – nothing more, nothing less. Whether or not the sets can be mapped to a phase space together or whether or not they lead to a repeated factor of some sort like a *strange attractor* or a *fractal coefficient* is inconsequential. Secondly, and more important, the measure of mutual information tells me on a visceral level whether or not a series of measurements is like or unlike another relative to a third element. It allows us to determine if the two measurements are indeed purely random, thus contain no mutual information relative to the third element, or are alike.

## III. Theoretical Underpinnings

We include a brief primer here on all three models so the reader can use this as a primer.

### A. Markov Chains

In order to consider a student progressing on a random walk, we'll need to formulate some introductory theory. If we define $X_n$ as the set of possible states in time for a random variable $X, n \in \mathbb{N}$ describing the various epochs in which the random variable is defined. Suppose $X_n$ takes values in the state space $S = \{1, 2, \dots, M\}$ and that the epochs over which it is defined is a discrete time space. To define a Markov chain we define:

$$P(X_n = i_n | X_1 = i_1, \dots, X_{n-1} = i_{n-1}) \, , i_k \in S \, \forall \, k = 1, 2, \dots \quad (1)$$

We then define the *Markov Property* [24] which reduces the probability in (1) to obtain,

$$P(X_n = i_n | X_1 = i_1, \dots, X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1}) \, , i_k \in S \, \forall k = 1, 2, \dots \quad (2)$$

This equation means in common terms that the previous past measurements are *conditionally* independent from the future given the most recent calculations. This is the most powerful assumption in Markov chain analysis.

The second assumption we exploit is *time homogeneity*, which allows us to assume that probabilities that define state transitions do not change over time and leads to the following,

$$P(X_n = i_n | X_{n-1} = i_{n-1}) = p(i_{n-1}, i_n) \quad (3)$$

Finally, we build what is called the probability transition matrix. We denote this matrix with $P$, and allow the $ij^{th}$ entry of the matrix to be the probability of transitioning from state $i$ to state $j$ for all $i, j \in S$. We also normalize, so that the sum of the rows of the matrix add up to one. This is because each row refers to the current state of our Markov system and the each column refers to an immediate future state.

$$P_{ij} = p(i, j) \, ; i, j \in S \quad (4)$$

$$\sum_j P_{ij} = 1 \, \forall i, j \in S \quad (5)$$

### B. Hierarchical Linear Model

All HLMs/MLMs begin with the construction of the linear regression formulae,

$$Y_i = \beta_0 + \beta_1 X_{i,} + e_i \quad (6)$$

$Y_i$ is the dependent variable, $\beta_0$ is the intercept, $\beta_1$ is the slope, $X_i$ is the predictor variable, and $e_i$ is the residual. Depending on the level and treatment of the equation itself, the residual is also the *error*.

In HLM, a standard model equation looks like,

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_i, + e_{ij} \quad (7)$$

Here $j$ here denotes our clustered unit factor. For every subset of $j$, there is a subset that includes an outcome (dependent variable), an intercept, a residual, and a slope.

For the papers that published with MIDFIELD data so far, the following equation has been exploited repeatedly,

$$Y_{ij} = \beta_{0j} + e_{ij} \quad (8)$$

This is called the *null model* or the *intercept only model*, and in some disciplines the *empty model* or the *fully unconditional model* [12].

$$ICC = \rho = \frac{\tau_{00}}{\tau_{00}+\sigma^2} = \frac{UN(1,1)}{UN(1,1)+e_{ij}} \quad (9)$$

Finally, we calculate the *intra class correlation* factor. The factor $\tau_{00}$ is the variability between levels and $\sigma^2$ is the variability within levels. To interpret this, here is an example, a value of the ICC of 0.10 means that 10% of the variability in an outcome lies within the nested or hierarchical structure we have defined. If we have defined the levels as being different classrooms the students occupy, then that means 10% of the variability lies within that structure.

### C. Mutual Information Theory

Mutual information theory depends on the extrapolation of the following equation, which should be familiar to those from an educational statistics background or who have taken statistical mechanics,

$$I(X;Y) \coloneqq \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = H(X) - H(X|Y) \quad (7)$$

Breaking this master equation down into component parts, the most basic is the familiar *Shannon Entropy*, defined by [25], the entropy of a random variable can be expressed by the following quantity,

$$H(X) = -\sum_i p_i \log(p_i) \quad (8)$$

By using a base of 2, we can create a "dimensionless" unit of measure – referred to as *bits*. The value $p_i$ then is then defined as the probability of event $i$ given that $i$ is located within the event space of the random variable. In other word, we now have a way of measuring the uncertainty around of the random variable $X$.

## IV. Preliminary Results of All Three Methods

All three methods have produced novel results in MIDFIELD using preliminary data, and we will present them here. The Markov results study the relationship between two state spaces – the terms spend within the university system and the number of major changes in the system notwithstanding switching out of *first-year engineering*. The HLM method studies the amount of variance of the final grade related to the overall *course size* of first year classes common among engineers, and the amount of grade variance endemic to the *section* of a course the student is in of a first year course.

### A. Markov Chain Results

The with Markov chains allow one to consider a *random walk* of students along a various progression. We care primarily about two factors: first, the number of terms enrolled at the university; and second, the number of changes of degree program (or major). If students who have excessively switched programs are removed (stopping at over twenty – yes, such student exist). For brevity, consider one *set* of engineering students, all aggregated students in the MIDFIELD data set. In the larger paper for this work, students who switched majors multiple times and were *only* ever engineers were considered, as well as students who switched multiple times and had engineering as their final declared major.

For leaving or graduating the system for the set of *all* students, there are two states derived from an analysis of a transfer matrix of our Markov chain: first, a student who fails to graduate or F; and second, a student who graduates or G. Summarizing the analysis for the first four major changes for the set of all students, the following can be derived,

| Major Switch # | F | G |
|---|---|---|
| 0 | 0.51695 | 0.483051 |
| 1 | 0.387187 | 0.612813 |
| 2 | 0.330237 | 0.669763 |
| 3 | 0.304362 | 0.695638 |
| 4 | 0.29814 | 0.70186 |

Table 1. Failure to graduate (F) versus graduation (G)

### B. Hierarchical Linear Model Results

For this cut of the MIDFIELD, there are 701,190 first-time-in-college students matriculating in any major at participating institutions [7, 8]. The first cut of data began with 137,071 first-time-in-college (FTIC) students who ever matriculated in engineering at one of nine MIDFIELD institutions between 1988 and 2011. The data for these students is *complete*, in other words we have everything that the registrar has for these students. There were 161,456 grades for instances of where students who sometime in their careers declared an engineering

major took one of three introductory courses: chemistry I; calculus I; and physics I.

From this analyzes, two sets of charts have been produced. In Table 2, we have the ICC numbers for a HLM that analyzes the grade variance endemic to the *section* a student is in of a set of calculus I courses across multiple institutions. Table 3 demonstrates the grade variance contained by the course *size* of all calculus I courses across multiple institutions in the MIDFIELD database.

| Institution | UN(1,1) | SE | Residual | SE | ICC | Intercept | SE |
|---|---|---|---|---|---|---|---|
| ALL | 0.2507 | 0.0089 | 1.4429 | 0.0083 | 0.1480 | 2.3047 | 0.0099 |
| 1 | 0.1903 | 0.0161 | 1.5184 | 0.0230 | 0.1114 | 2.5309 | 0.0184 |
| 2 | 0.1096 | 0.0252 | 1.6834 | 0.0469 | 0.0611 | 1.9803 | 0.0299 |
| 3 | 0.1958 | 0.0184 | 1.5044 | 0.0235 | 0.1152 | 2.1321 | 0.0237 |
| 4 | 0.155 | 0.0152 | 1.3719 | 0.0202 | 0.1015 | 2.4528 | 0.0227 |
| 5 | 0.1695 | 0.0177 | 1.3168 | 0.0159 | 0.1140 | 2.5186 | 0.0233 |
| 6 | 0.3222 | 0.0431 | 1.6478 | 0.0482 | 0.1636 | 2.3323 | 0.0403 |
| 7 | 0.0473 | 0.0086 | 1.2952 | 0.0189 | 0.0352 | 2.3486 | 0.0229 |
| 8 | 0.4001 | 0.0524 | 1.7531 | 0.0434 | 0.1858 | 2.0184 | 0.0482 |
| 9 | 0.1743 | 0.0292 | 1.5974 | 0.0436 | 0.0984 | 1.8378 | 0.0332 |

Table 2. Table of results from the first core calculus course

| Institution | UN(1,1) | SE | Residual | SE | ICC | Intercept | SE |
|---|---|---|---|---|---|---|---|
| ALL | 0.0662 | 0.0081 | 1.2416 | 0.0079 | 0.0506 | 2.3047 | 0.0099 |
| 1 | 0.1064 | 0.0298 | 1.6443 | 0.0250 | 0.0608 | 2.5309 | 0.0184 |
| 2 | 0.0953 | 0.0363 | 1.9132 | 0.0526 | 0.0474 | 1.9803 | 0.0299 |
| 3 | 0.0159 | 0.0101 | 1.7758 | 0.0464 | 0.0089 | 2.1321 | 0.0237 |
| 4 | 0.0128 | 0.0056 | 1.5109 | 0.0218 | 0.0084 | 2.4528 | 0.0227 |
| 5 | 0.1068 | 0.0321 | 2.0088 | 0.0483 | 0.0505 | 2.5186 | 0.0233 |
| 6 | 0.0238 | 0.0085 | 1.5395 | 0.0255 | 0.0153 | 2.3323 | 0.0403 |
| 7 | 0.0336 | 0.0083 | 1.3098 | 0.0191 | 0.0250 | 2.3486 | 0.0229 |
| 8 | 0.0634 | 0.0228 | 1.6661 | 0.0239 | 0.0366 | 2.0184 | 0.0482 |
| 9 | 0.0587 | 0.0242 | 1.7281 | 0.0447 | 0.0329 | 1.8378 | 0.0332 |

Table 3. Table of results from the first core calculus course

### *C. Mutual Information Results*

The mutual information treatment is perhaps the simplest of the three. For the same set of data as in the Markov analysis, the mutual information treatment assumes a random variable of *S,* which represents the *number of times a student switches majors*, and two subsequent variables, G and T, which represent the student's *Final GPA* and the *Final Term Enrolled*, respectively. For the total cohort of students, we have three sets, *all of the students*, *those who leave*, and *those who graduated.*

| All | | Leaving | | Grad | |
|---|---|---|---|---|---|
| I(T;S) | 0.1111 | I(T;S) | 0.1409 | I(T;S) | 0.0284 |
| I(G;S) | 0.0317 | I(G;S) | 0.0312 | I(G;S) | 0.0187 |

Table 4. A randomness relationship between variables

## V. Discussion and Conclusion

Among the Markov chain models that have been produced, one novel relationship is that they match the conclusions of the aggregate study of the effect of switching majors multiple times performed on the complete MIDFIELD data set [2]. One parallel result that is impressive is that both the aggregate data and the Markov analysis conclude that for students who remain after switching a major, the probability of success (aka graduation) increases per successful change. While this may not be intuitive, it is almost universal across MIDFIELD. In other words, "Those who survive the switch end up having a higher chance of graduating per switch."

The hierarchical linear model conclusions are perhaps more important and novel. First, the effect of course *size* being indeterminate to the point where a structural equation model (SEM) may be more useful is important. Generally, when there is less than 3% variance explained by the presence of a nested structure, an SEM treatment will prove valuable. This is one area that MIDFIELD has yet to touch and would prove complimentary to these results. Second, the HLM analysis of course *section* has demonstrated that for three sets of introductory courses most engineers take at MIDFIELD institutions, the section of the course is in on any one semester will determine to a great degree the *variance* of the final grade! While this is *not enough* to conclude whether or not taking one course over another yields a *higher grade*, an industrious student with a little bit of institutional data could determine the *grand mean* (aka overall average) of grades for any one course, and from that make a good prediction whether or not switching sections was worth it given this data.

The mutual information results are straightforward. If we consider *all* students in the database, we find that between the final GPA of all students and the final term enrolled of all students, there is a separating effect between those who have graduated and those who have not (or have simply left the database). Another way of saying this is that the *number of times one switches majors* is seemingly more random for those who leave the institution or fail to graduate than ones who do graduate. Once again, this is another item that seems non-intuitive, but makes sense. Perhaps students who switch majors who graduate have done some with a greater deterministic purpose, as opposed to at random.

## References


[1] G. Ricco, N. Salzman, R. Long, and M. Ohland, "Sectionality or Why Section Determines Grades: an Exploration of Engineering Core Course Section Grades using a Hierarchical Linear Model and the Multiple-Institution Database for Investigating Engineering Longitudinal Development," in *American Society for Engineering Education Annual Conference*, 2012.

[2] G. Ricco, "Degree Program Changes and Curricular Flexibility: Addressing Long Held Beliefs About Student Progression," PhD, Purdue University, West Lafayette, IN, 2013.

[3] G. Ricco, "How Course Size Effects Grades: Sizeness and the Exploration of the Multiple - Institution Database for Investigating Engineering Longitudinal Development through Hierarchal Linear Models," presented at the American Society for Engineering Education, 2015.

[4] G. Ricco and J. Ryan, "Major Changes and Attrition: An Information Theoretic and Statistical Examination of Cohort Features Stratified on Major Switches," presented at the American Society for Engineering Education, 2015.

[5] G. D. Ricco and M. Hammond, "Exploration of Degree Program Change: A Novel Use of Nearest Neighbor Classifiers," presented at the American Society for Engineering Education, 2020.

[6] G. D. Ricco and M. W. Ohland, "A Survival Analysis Model of Students Who Switch Degree Programs: A Novel Treatment of Student Attrition," 2020.

[7] R. A. Long. "The Multiple-Institution Database for Investigating Engineering Longitudinal Development." https://engineering.purdue.edu/MIDFIELD (accessed.

[8] M. W. Ohland, S. D. Sheppard, G. Lichtenstein, O. Eris, D. Chachra, and R. A. Layton, "Persistence, engagement, and migration in engineering programs," *Journal of Engineering Education,* vol. 97, no. Compendex, pp. 259-278, 2008.

[9] G. D. Ricco and J. F. Ryan, "A Random Walk on the Major Path Space: Examining Student Progression as a Random Process Using Markov Chains," 2020.

[10] K. Kaplan and J. Kaplan, "Markov Chains: Reintroducing Lost Knowledge Back into a Modeling and Simulation Course," in *Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition*, 2005.

[11] J. N. Johnstone and H. Philp, "The application of a Markov Chain in educational planning," *Socio-Economic Planning Sciences,* vol. 7, no. 3, pp. 283-294, 6// 1973, doi: http://dx.doi.org/10.1016/0038-0121(73)90020-7.

[12] S. Raudenbush and A. S. Bryk, "A Hierarchical Model for Studying School Effects," (in English), *Sociology of education,* Article vol. 59, no. 1, pp. 1-17, Jan 1986, doi: 10.2307/2112482.

[13] S. R. Porter, "Institutional structures and student engagement," (in English), *Research in Higher Education,* Article vol. 47, no. 5, pp. 521-558, Aug 2006, doi: 10.1007/s11162-005-9006-z.

[14] R. T. Jewell, M. A. McPherson, and M. A. Tieslau, "Whose fault is it? Assigning blame for grade inflation in higher education," *Applied Economics,* vol. 45, no. 9, pp. 1185-1200, 2013/03/01 2011, doi: 10.1080/00036846.2011.621884.

[15] G. Brix, S. Zwick, F. Kiessling, and J. Griebel, "Pharmacokinetic analysis of tissue microcirculation using nested models: Multimodel inference and parameter identifiability," *Medical Physics,* vol. 36, no. 7, pp. 2923-2933, Jul 2009, doi: 10.1118/1.3147145.

[16] R. Cudeck and J. R. Harring, "Analysis of nonlinear patterns of change with random coefficient models," in *Annual Review of Psychology*, vol. 58, (Annual Review of Psychology, 2007, pp. 615-637.

[17] G. Brostrom and H. Holmberg, "Generalized linear models with clustered data: Fixed and random effects models," *Computational Statistics & Data Analysis,* vol. 55, no. 12, pp. 3123-3134, Dec 2011, doi: 10.1016/j.csda.2011.06.011.

[18] R. R. Dinu and A. Veeraragavan, "Random parameter models for accident prediction on two-lane undivided highways in India," *Journal of Safety Research,* vol. 42, no. 1, pp. 39-42, Feb 2011, doi: 10.1016/j.jsr.2010.11.007.

[19] A. S. A. Mohamed and T. Prasad, "Analysis of the neuromuscular system using random parameter models," *Mathematical and Computer Modelling,* vol. 13, no. 4, pp. 19-36, 1990, doi: 10.1016/0895-7177(90)90050-w.

[20] R. Craine and A. Havenner, "Classical versus Bayesian models-on the dangers of a little bit of knowledge," *International Journal of Systems Science,* vol. 14, no. 8, pp. 871-5, 1983.

[21] C. Rodrigues Neto and A. C. R. Martins, "Multifractality in the random parameter model for multivariate time series," *Physica A: Statistical Mechanics and its Applications,* vol. 388, no. 11, pp. 2198-2206, 2009, doi: 10.1016/j.physa.2009.02.005.

[22] N. C. f. E. Statistics, "High School and Beyond: Sample Design Report," Washington D.C., 1980.

[23] J. Coleman, T. Hoffer, and S. Kilgore, "Cognitive Outcomes in Public and Private Schools," *Sociology of education,* vol. 55, no. 2/3, pp. 65-76, 1982. [Online]. Available: http://www.jstor.org/stable/2112288.

[24] G. Lawler, *Introduction to Stochastic Processes*. New York City: Chapman and Hall, 1995.

[25] C. Shannon, "A Mathematical Theory of Communication " *The Bell Systems Technical Journal,* vol. 27, 1948.