Leveraging Access to and Use of Department of Defense Data: A Case Study of Unraveling Military Attrition Through New Approaches to DoD Data Integration

FCSM Research and Policy Conference

March 8, 2018

Joshua Goldstein

VirginiaTech. Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Authors

US Army Research Institute for the Behavioral & Social Sciences

- Andrew Slaughter, Senior Research Psychologist

Social and Decision Analytics Laboratory,
Biocomplexity Institute

- David Higdon, Professor of Statistics
- Sallie Keller, Professor of Statistics and Director
- Stephanie Shipp, Research Professor and Deputy Director
- Vicki Lancaster, Senior Research Scientist
- Bianica Pires, Research Scientist

Biocomplexity Institute of Virginia Tech

Problem-Driven Science

Our information biology approach is putting research to work in the real world, breaking down barriers between science and policy.

Social and Decision Analytics Laboratory

The Social and Decision Analytics Laboratory brings together statisticians and social and behavioral scientists to embrace today's data revolution, developing evidence-based research and quantitative methods to inform policy decision-making.

Motivation: The Challenge

"Our people are the most significant weapon in our arsenal"

- Half of the U.S. Defense Budget is spent on these people, e.g., pay, benefits, health care, training, housing, child care, etc.

Should *all data* be leveraged into these decisions?

- The Army population is highly dynamic and must be analyzed within context and on a longitudinal scale
- Need to gain a socio-demographic understanding
- Need integrated eco-system of multiple data sources

DoD ARI / VT SDAL Collaborative Research

Research in science of ALL data to support DoD

- Assess ability to access, evaluate the quality, and integrate DoD data and other data to support decision making related to the ***Army Population***
- Open opportunities to expand the types of questions and that can support findings beyond operational purposes
- Ground research in the context of a real problem
  - –Identify predictors of attrition within enlisted ranks

Research Questions

- What is the ***value of combining*** DoD, civilian, and non-federally collected ***data sources*** to enhance or complement a representative use of DMDC (PDE) data?
- How does this help capture and model individual, unit, and organizational characteristics and non-military ***contexts*** that affect attrition?
- How do we ***quantify this value***?

Research Objective

- To assess the best approaches for successfully accessing and integrating DoD data and other data sources within the context of a case study on military attrition of Army enlistees

Data Discovery and Profiling in the Person Data Environment

VirginiaTech.
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Person Data Environment (PDE)

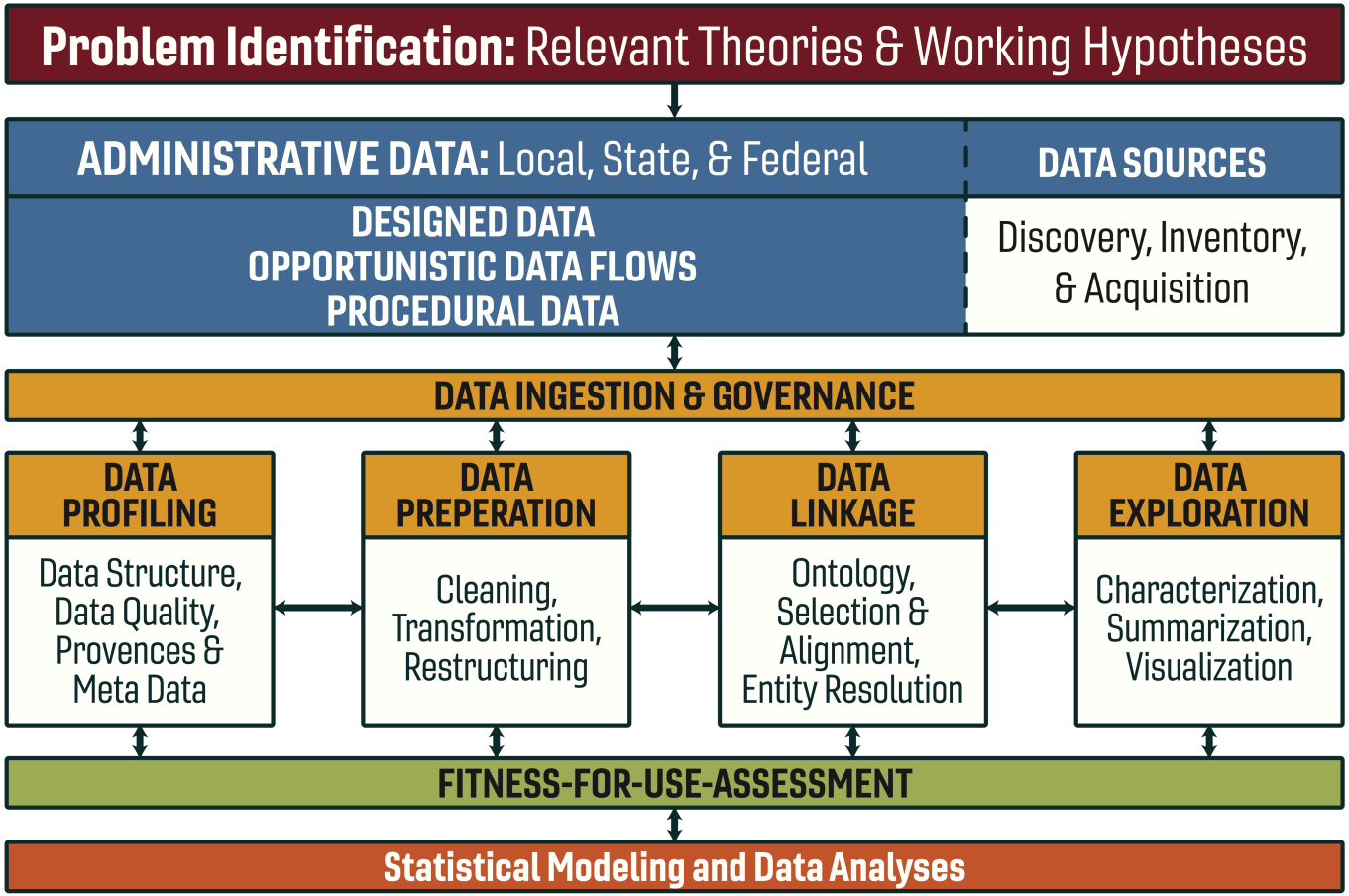
- DoD maintains numerous datasets about military personnel and their families, including their military service information including deployments, demographics, accessions, pay, and waivers, as well as their health, training, and positions held.
- DMDC provides access to many of these datasets through the Person Data Environment (PDE) in a secure, cloud-based enclave.

PDE Access

To access the data requires the completion of necessary administrative steps, many which initially seem as daunting as the actual research. These involve:

1. Obtaining CAC (Common Access Card) cards for non-military academics,
2. Completing the Institutional Review Board (IRB)/Human Research Protections Official (HRPO) processes; and
3. Working through the mechanics of accessing the PDE/CITRIX system.

Data Science Framework



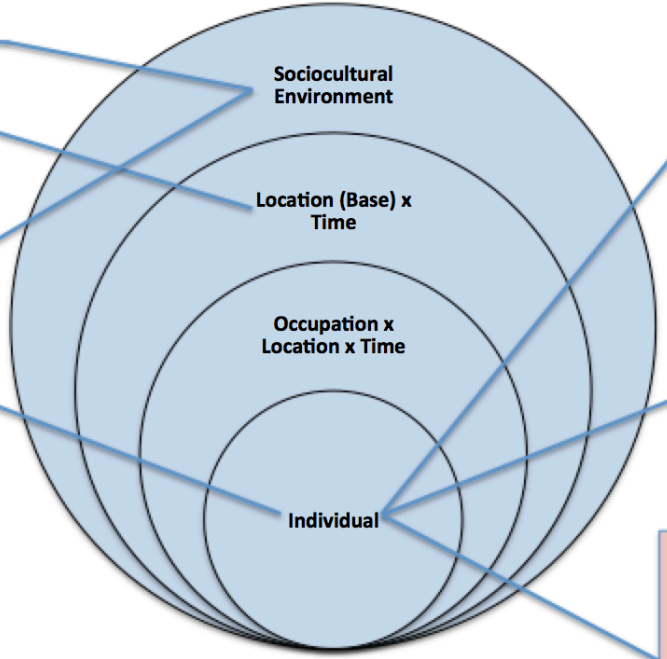
Towards an Integrated Data Framework for Understanding the Context of Army Environments

Other Information

Variables We Have
- Policy changes (e.g., peacetime vs. war)
- Non-personal shock events (e.g., 9/11)
- Job alternatives (e.g., ACS employment)
- Local community (e.g., ACS data)

Constructs to be Modeled
- National Army prestige/support
- Cohesion
- Job satisfaction
- Job investment
- Commitment norms

Data and Information Map*



Administrative Data

Demographics
- Race
- Ethnicity
- Sex
- Birthdate/Age
- Faith group
- Education level and discipline
- Marital status
- Spouse in military indicator
- Number and type of dependents
- ASVAB score
- State/country of residence before entry

Service Dates and Locations
- Length of time in service
- Length of service agreement
- Location (base) over time
- Obligation begin and end dates
- Term of service
- Date of initial entry
- Date of end of initial training

Military-Specific Characteristics/Incentives
- Security clearance
- Education incentive indicator
- Career status bonus program indicator
- Object of mission (e.g., advanced cruise missile)
- Occupation group (primary and secondary)
- Re-enlistment eligibility
- Aeronautical rating code (e.g., astronaut)
- Flying status indicator
- Pay grade (e.g., E-3) and length of time in grade
- Character of service (e.g., honorable)

To understand the individual, it is necessary to understand the social contexts in which the individual is embedded. Social contexts represent an ecological system composed of different levels, ranging from the immediate environment to higher social and cultural forces.

*Unit level not available to us.

Data Profiling

- Goal: Determine quality and utility of data for research use
- Looking at measures such as completeness, value validity, longitudinal consistency, and uniqueness.
- Example problems include changes in an individual's recorded gender over time, duplication of records by ID, etc.
- Develop a ‘gold standard’ for demographics profiling

Steps in the Data Quality Analysis Process

Completeness	**The Proportion of Elements Properly Populated for a Given Purpose** • Issue types include: record fields containing no data; records not containing necessary fields; datasets not containing the requisite records (e.g., Testing for NULLs and empty strings existing where not appropriate)
Validity	**The Proportion of Elements whose Attributes Possess Proper Values** • Checking for value validity generally comes in the form of straight-forward domain constraint rules • e.g., count gender where gender is not (male, female), count age where age is not between [0, 110]
Uniqueness	**The Count of Unique Values Taken by an Attribute or Combination of Attributes** • Frequency distribution of an element. (Note: The more homogeneous the data values of an element, the less useful the element is for analysis)
Duplication	**The Degree of Replication of Distinct Observations Per Observation Unit Type** • e.g. Greater than 1 registration per student per official reporting period. • Note: Duplication occurs as a result of choice of level of aggregation
Consistency (Record)	**The Degree to which Two or More Data Attributes Satisfy a Well-Defined Dependency Constraint** • a.k.a. - relationship validation • e.g. zip-code – state consistency e.g. gender – pregnancy consistency
Consistency (Longitudinal)	**The Degree to which an Attribute, or Combination of Attributes, Remains Consistent Over Time** • e.g. An individual's gender changing and then changing back (clerical error) • e.g. An individual's race classification changing from one to two-races (change in number of options)

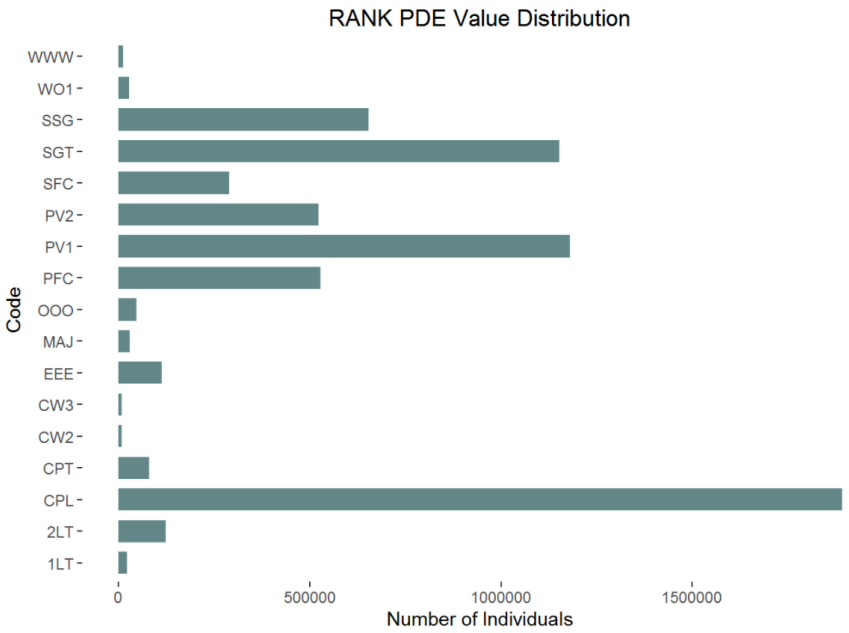
Data Profiling Example

SUMMARY OF DATA ELEMENT

DE_BUS_NM: Rank, PDE-Based DE_DEFN_TX: Servicemember's raw rank, obfuscated for PDE usage. Added by AAG ETL staff during the staging process. DE_DOMN_DSCR_TX: Indicates a servicemember's rank as obfuscated for PDE.

Number of Unique Values: 61

VALUE	VALUE_DESCRIPTION
1LT	First Lieutenant
1st Lt	First Lieutenant
1stLt	First Lieutenant
2LT	Second Lieutenant
2d Lt	Second Lietenant
2ndLt	Second Lieutenant
A1C	Airman First Class
AA	Airman Apprentice
AB	Airman Basic
AN	Airman
AR	Airman Recruit
Amn	Airman



DATA ANALYTICS SUMMARY OF RESULTS

Test	Measurementvalues	Value
Completeness	Number of missing values	756
	Percent of complete values	99.98 %
Validity	Number of invalid values	0
	Percent of valid values	100 %
Uniqueness	Number of unique values	17

Statistical demographic log analysis & reduction

Basic demographics between data files needed to be reconciled and reusable data products created.

- **Demographics Table**
  - Information about the enlistee that typically remains static over time, e.g., gender, race, ethnicity, entry test scores
  - Simple rules are applied to resolve duplicates and invalid entries
  - Contains one row per PID
- **Transaction Table**
  - Events or enlistee information that can change periodically, e.g., duty station, rank, pay grade
  - Contains multiple rows per PID

Column Name	Description	Original Table
PID_PDE	Enlistee's Unique ID	Master
PN_SEX_CD	Gender	Master
RACE_CD	Race Code	Master
INIT_ENT_TRN_END_DT	Initial Entry Training End Date	Master
DATE_BIRTH_PDE	Person Birth Date	Master
PN_BIRTH_PLC_CTRY_CD	Person Birth Place Country Code	Master
HOR_ZIP_CODE_PDE	Home of Record Zip Code	Analyst
ACT_SCORE	ACT Score	Analyst
SAT_SCORE	SAT Score	Analyst
AP	ASVAB: Auditory Perception Score	Analyst
CO	ASVAB: Combat Score	Analyst
.	.	.
.	.	.

Data Discovery - DoD Data Sources

Currently Accessible in PDE

- Active Duty Military Personnel Master
- Active Duty Military Personnel Transaction
- MEPCOM Regular Army Analyst
- Interactive Personnel Elective Records Management System (IPERMS)
- Digital Training Management System (DTMS)
- Army Career and Alumni Program (ACAP)
- Global Assessment Tool (GAT)
- Pre-Deployment, Post-Deployment, Periodic Health Assessments
- Unit Risk Inventory (URI)
- DEOMI Organization Climate Survey
- Omaha 5 Behavioral Survey

Data Discovery – Non-DoD Data Sources

Data external to the PDE provide richer information on the surrounding social, political, and economic environments that affect an individual's decision to attrit.

- Tables ingested to by county and year (2005-2015) are available to all researchers in the PDE.
- Crosswalk of counties to deployed installation, home county.

Data sources:

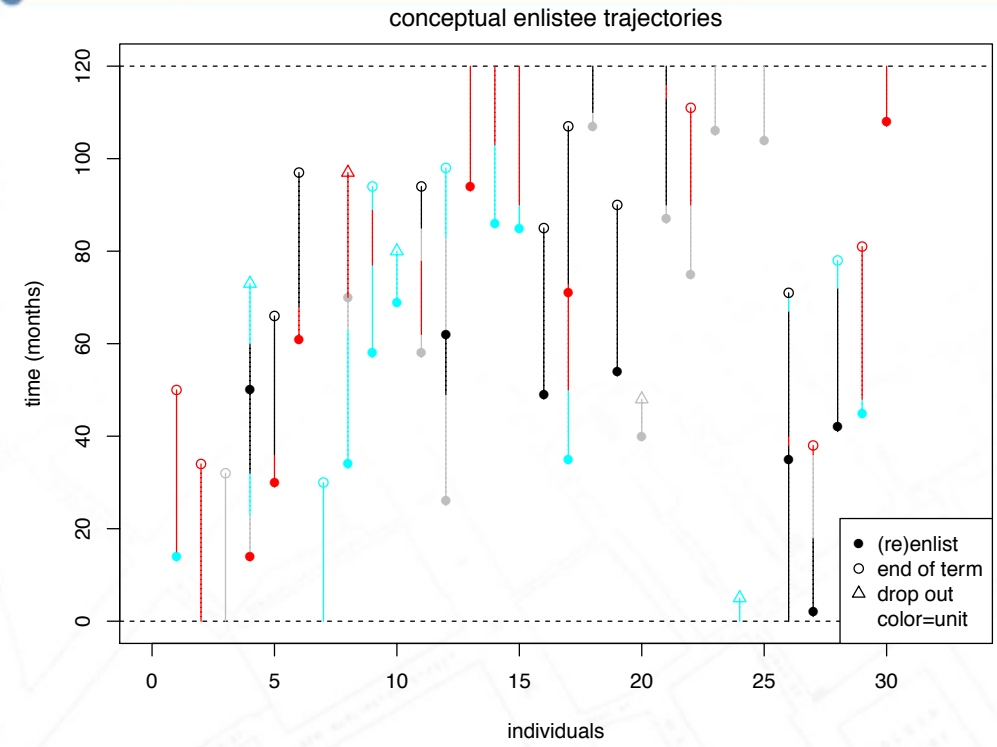
- American Community Survey (ACS) from the U.S. Census Bureau
- Bureau of Labor Statistics data including
  - Occupational Employments Statistics
  - Quarterly Census of Employment and Wages (QCEW)
  - Current Employment Survey (CES) data

Case Study: Modeling First-Term Army Attrition

VirginiaTech
Biocomplexity Institute

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Conceptual Model of Attrition



- Enlistees through time may transfer units, reenlist, and attrit.

VirginiaTech
Biocomplexity Institute

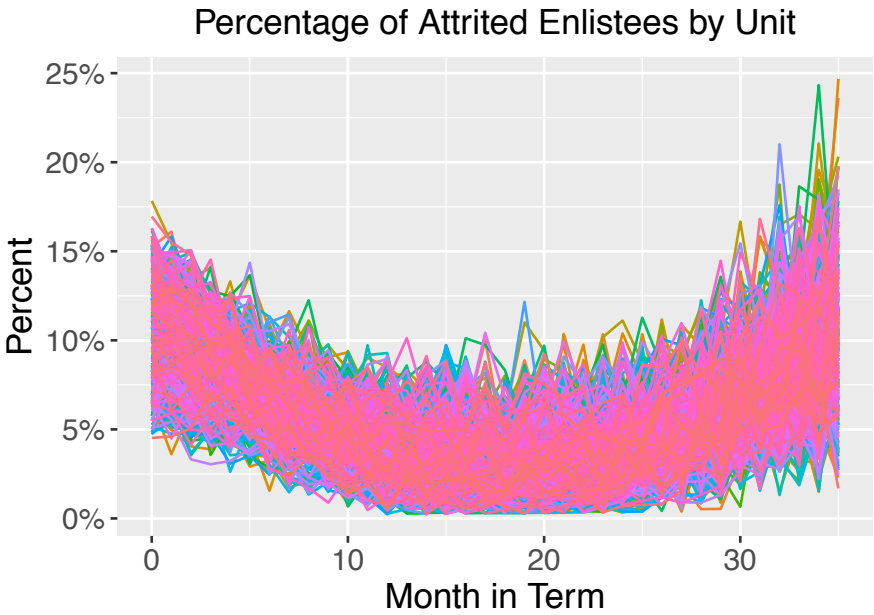
An Agent-based Model of Attrition

Develop an agent-based model for the purpose of generating simulated data to test potential statistical models.

The ABM simulates the enlistment, attrition, and re-enlistment of active enlistees.

- Each agent is assigned to a Unit; attrition probability follows a U-shaped curve
- For each simulation month tick, agents can perform one of the following actions: enlist, transfer units, re-enlist, attrit.

Simulation output provides synthetic data of enlistee attrition behavior across time.



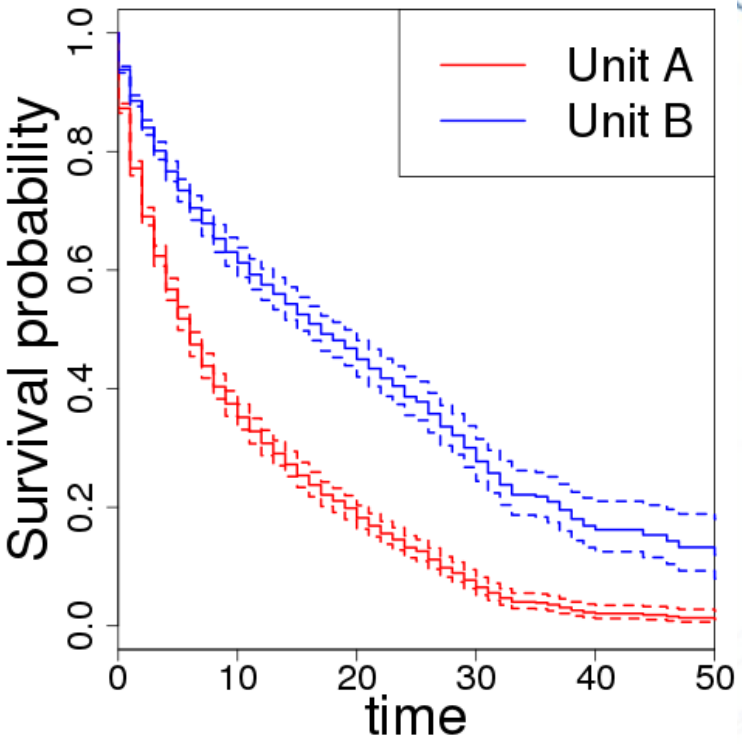
Proportional Hazards Model

Hazard rate at time *t* for agent *i*

$$\lambda_i(t) = \lambda_0(t)\exp(X_i\beta)$$

where X_i is a vector of covariates.

For the simulated example (right), X_i is the Soldier's unit; fitted survival probabilities show an effect of unit on attrition.



VirginiaTech.
Biocomplexity Institute

Building Complexity

- Model flexibility for connecting many data sources
- Need to integrate "external" data sources that change over time.
- Need to integrate person-specific information
  - Relevant time is with respect to a Soldier's term
  - "Exposures" to duties, leaders, training, …
- Priors needed to estimate many similar effects (e.g. unit, duty location)

Multinomial Proportional Odds

Cox discrete proportional odds model (Cox 1972); discrete-time version of proportional hazards model with time-varying covariates $X(t)$.

Extended to allow for competing risks for multiple events; *r* denotes the risk (Allison 1982).

$$\lambda_i^r(t) = \frac{\exp(\,\beta_0^r(t) + \beta^r X_i(t)\,)}{1 + \sum_m \exp(\,\beta_0^m(t) + \beta^m X_i(t)\,)}$$

- Define $\lambda_i^r(t)$ as the cause-specific hazard, i.e. the probability that an attrition of Soldier *i* occurs due to risk *r* at time *t*, given that the event had not occurred at *t-1*.
- Time-varying data represented in counting process format
- Hazard related to covariates as in multinomial logistic regression.

VirginiaTech
Biocomplexity Institute

Bayesian MPO Model

Bayesian discrete time, additive model for the rate of Soldier attrition. This model builds on the multivariate proportional odds model (King 2014).

$\lambda_i^r(t)$ is the cause specific hazard, $\lambda_i^0(t)$ is the probability of no event.

$$\eta_i^r(t) = \log\left(\frac{\lambda_i^r(t)}{\lambda_i^0(t)}\right)$$

$$\eta_i^r(t) = \beta_0^r(t) + \sum_k \beta_k^r(X_{k,i}(t))$$

- If $X_{k,i}(t)$ is categorical, estimate a separate effect for each category.
- For continuous covariates, $\beta_k^r(X_{k,i}(t))$ is piecewise constant with a Gaussian Markov Random Field prior.
- For a random effect, $(\beta_k^1, \dots, \beta_k^R) \sim N_R(0, \Sigma_k)$ where Σ_k has an inverse Wishart specification.

Model Features

This is a flexible hierarchical modeling framework that can account for:

- A random effect for individual Soldiers (e.g. 'taste' for the military)
- Nonlinear, time-varying covariate effects
- Competing risks due to multiple causes of attrition
- Recurrent events due to reenlistment

Efficient Bayesian inference for model parameters via Markov chain Monte Carlo (MCMC).

Model Covariates

Covariates observed quarterly for >1M soldiers, 2005-2015. They include information on

- Demographics (Age, Gender, Race, Education, Marital Status)
- Military Characteristics (Rank, MOS, Unit, Court Martials, Deployments)
- Testing (ASVAB, APFT, Training)
- Community Factors (Employment rate, Wages, Poverty, etc.)

Conclusion

We integrated multiple data sources within and external to the DoD, and levereaged this data for a case study on Soldier attrition. Next steps:

- Exploring attrition by MOS, unit, geography.
- Predictive power of our model and comparison to baselines.

References

- Cox, D. R. (1972). *Regression models and life-tables.* Journal of the Royal Statistical Society, Series B, 34(2):187–220.
- Allison, P. D. (1982). *Discrete-time methods for the analysis of event histories.* Sociological Methodology, 13(1):61–98.
- King, A.J. (2014). *Bayesian Event History Analysis with Applications to Recurrent Episodes of Illicit Drug Use.* Ph.D. Thesis, University of California, Los Angeles.