tbanalytics99@gmail.com **Thomas J. Balmat** 919-820-1057

Research Data Analyst and Statistician, Duke University, Durham, North Carolina
Concentration in statistical modeling and computation involving large, complex data sets

Experience

Jul 2018 - Present **Research Analyst and Computational Specialist**
Research Computing, Duke University, Durham, North Carolina

The Research Computing Department at Duke offers a high performance computing cluster (the Duke Computing Cluster, or DCC https://oit-rc.pages.oit.duke.edu/rcsupportdocs/about/) consisting of 1,360 compute nodes, more than 45,000 virtual CPUs, 980 GPUs, and 270TB of RAM. Faculty, post-docs, and graduate students use the DCC for intensive computational aspects of advanced research. My primary role in DCC support is to assist researchers in improving computational efficiency, specifically with R (statistical software). This typically involves various parallel processing techniques, including use of the parallel package, the Rcpp package (combined with OpenMP), NVIDIA Cuda programming (for GPUs), and Slurm scripting for CPU and memory resourcing.

While supporting researchers with their R needs on the DCC, I have maintained involvement in various research projects in the Law School, Fuqua Business School, School of Nursing, Pratt School of Engineering, Department of Molecular Genetics and Microbiology, Nicholas School of the Environment, Statistical Sciences, and other departments. Example projects include:

- Hi-Host Phenome (H2P2) Project (http://h2p2.oit.duke.edu). H2P2 is a genome wide association study (GWAS) that identifies significant correlative relationships between pathogen induced human cell damage and cell genotype. The study involves approximately 150 pathogens, 15 million genotypes, and 500 human subjects. A SQL database consisting of more than 12 billion observations was assembled and parallelized performance strategies were implemented to achieve efficient query results.
- Interactive Cross Phenotype Analysis of GWAS Database (iCPAGdb, http://cpag.oit.duke.edu). iCPAGdb is an R/Shiny application that compares significant traits (phenotypes), by common genetic classification, from independent GWAS studies. Researchers can upload a GWAS data set and conduct cross-phenotype analysis in real-time. Challenges with this project include efficient computation of statistical results and presentation of informative graphics, given the dynamic range of input data sets.
- Human Capital Synthetic Data Project. The Synthetic Data Project models demographic, skill, pay, and advancement characteristics for U.S. federal employees using career longitudinal and covariate relationships for 3,500,000 employees over a twenty-four year period. Various statistical estimation and differential privacy methods were employed to accurately model covariate authentic relationships, while limiting the risk of revealing sensitive information on individual employees.

Please refer to https://github.com/tbalmat/CV/blob/master/TomBalmatRecentProjects.pdf for details on additional projects.

Jul 2015 – Jun 2018 **Research Data Analyst and Statistician**
Social Science Research Institute (SSRI) and Department of Statistical Science
Duke University, Durham, North Carolina

The Social Science Research Institute at Duke University (https://ssri.duke.edu/) is home to fourteen centers for research in the social sciences. I was involved in two projects: U.S. Federal Human Capital (SSRI) and Synthetic Social Science Data (Dept. of Statistical Science). The Human Capital project attempts to model

policy value and effectiveness by the career profiles and qualifications of associated leaders and employees in the U.S. government. It involves intensive data management and statistical model development using a large, historical database of federal employee career profiles, along with historical economic and policy characteristics. The objective of the Synthetic Data project is to develop an anonymous version of the authentic Human Capital database that maintains career longitudinal and covariate relationships while limiting risk of individual employee identification.

Significant activities:

- Design and implement a SQL database to house and secure over 500,000,000 observations of historical data related to U.S. federal workforce human capital, the economy, governmental policy, and various social indicators
- Import raw observations from the Office of Personnel Management, the Centers for Disease Control, the Department of Education, and other sources into hierarchical, relational SQL structures
- Design and implement SQL queries involving up to 300,000,000 observations for efficient research model construction in R and Stata
- Identify computational inefficiencies in statistical estimation algorithms implemented in R (regression, aggregation, linear algebraic) and redesign using parallel processing methods
- Develop efficient algorithms for solving large linear, non-linear, logistic, quantile, and fixed effects regression models involving designs of dimension up to 30,000,000 X 2,000 using X'X indicator-sum and sparse matrix methods
- Assist with model development (linear, logistic, and quantile regression) for published research in areas such as gender/race wage gaps, promotion disparity, and grade inflation in the U.S. federal government
- Verify the consistency of covariate relationships between synthetic and authentic data using joint density and mass distributions, correlation matrices (with interactions, dimensions reach 40,000 X 40,000), and a variety of comparative bar, kernel, and surface plots
- Verify the utility of synthetic data by comparing expected values and parameter estimates from actual research models fit to synthetic and authentic data; report with a variety of comparative t-test, residual, and $\beta_{synthetic}$ vs. $\beta_{authentic}$ plots, paneled by estimator category (gender, race, age, education, occupation)

Mar 2014 – Jun 2015 Laboratory Data Manager and Statistician, Cormetech Inc., Durham, North Carolina

Cormetech (www.cormetech.com) operates one of the largest environmental selective catalytic reduction (SCR) testing laboratories in the United States. The lab simulates coal and natural gas power plant emissions in thousands of physical and chemical reaction tests each month. The resulting data require significant organization, certification, and archival.

Significant activities:

- Development of a single source SQL based lab request system to replace a variety of free-form and untraced telephone, e-mail, and paper requests
- Development of a comprehensive SQL lab result database for test integrity rule enforcement, final result reporting, and catalyst performance statistical modeling
- Development of reactor and test equipment key parameter control limits using historical data, linear and non-linear regression, moment matching, population parameter modeling with simulated data subsetting
- Development of SQL based lab capacity and cost reports for revenue reconciliation and cost accounting
- Development of SQL interfaces to various commercial and engineering databases to eliminate inefficient data replication and errors due to obsolete customer design and operational information
- Instruct lab staff on proper use of probability distributions for hypothesis testing, sample size specification for type II error control, model challenge methods, and robust confidence interval development

tbanalytics99@gmail.com **Thomas J. Balmat** 919-820-1057

Jan 1998 – Feb 2014 Operations Analyst, Hanson Pipe and Precast, Dunn, North Carolina

Develop and implement solutions to improve customer service, operational efficiency, quality, and business intelligence in all critical areas, including strategic development, sales, customer service, scheduling, inventory management, dispatch management, and labor efficiency. Since 1998 the company, as a team, doubled its revenue and service capacity with no significant increase in inventory, backlog, or staff, while significantly improving efficiency, quality, and customer satisfaction.

Significant activities:

- Facilitation of study and improvement teams, development of objectives, strategy, and methods
- Development of optimal business transactions and methods, performance metric definition
- SQL data summarization for key indicator performance reporting, response/predictor data modeling
- Statistical modeling of operational effectiveness and efficiency - regression, ANOVA
- Design of experiments to explore new operational models and influential effects - factorials, regression
- Optimal decision modeling, process state and cost/benefit analysis - stochastic methods
- Throughput analysis, queue modeling, Poisson processes, service model and capacity analysis
- Graphical presentation methods, performance charts, automated periodic key indicator reports

Oct 1992 – Dec 1997 Quality Assurance Statistician, Morganite Incorporated, Dunn, North Carolina

Evaluate product quality and performance using data from manufacturing processes, simulated life-cycle laboratory experiments, and customer supplied field tests.

Significant activities:

- Develop statistical models to project defects per million from historical inspection data, extensive experience with non-normal distributions such as lognormal, exponential, and Johnson S-family
- Design experiments for “zero-defect” process improvement and significant factor identification, factorial/fractional factorial designs, response surface and optimal target development
- Develop multiple parameter process characterization models (regression), assess influential observations, residual normality, and model robustness, residual/zone dependencies
- Tolerance analysis and design, Taylor series approximation, non-linear programming, simulation
- Process control chart development and implementation, mean, variance, repeated observations
- Sampling plan specification, operational curve (OC) development and implementation, type II error analysis, hypergeometric, binomial, multinomial, and Poisson probability analysis

Aug 1984 - Sep 1992 Information Systems Analyst/Programmer, Morganite Incorporated, Dunn, NC

Develop information management solutions in support of engineering, quality assurance, and inventory functions. Enable and support organizational improvement through development of database and programming solutions to meet the needs of key decision analysts and staff.

Key activities:

- Facilitation of team oriented business requirements and needs analysis
- System analysis and information flow design to meet objective requirements
- Database development, programming, implementation, support, and training

Software expertise

- High Performance Computing environment, including Slurm and NVIDA Cuda (C for GPUs)
- R, SAS, Stata, MS SQL Server, Oracle SQL, MS Access, Visual Basic, C, MS Windows, Unix/Linux

tbanalytics99@gmail.com **Thomas J. Balmat** 919-820-1057

- R packages: RODBC, ggplot2, parallel/snow, Rcpp, matrix, sparseM, quantreg, quantmod
- R Packages developed: feXTXc (large fixed effects regression solution), syntheticDataDP (differential privacy verification server functions)

Education

- Undergraduate curriculum in Mathematics, University of Houston, NC State University
- Graduate curriculum in Statistics and Industrial Systems Engineering (Mathematical Programming and Probability), NC State University

Certifications: ASQ Quality Engineer, SAS Base Programmer
Military service: U.S. Air Force, four years, honorably discharged, final grade E-5