thomas.balmat@duke.edu **Thomas J. Balmat** 919-820-1057

Significant Projects

Development of a logistic regression promotion model to estimate probability of promotion within the U.S. federal workforce, with emphasis on gender and race equality. Using approximately 35,000,000 records of career profiles of approximately 4,000,000 federal employees spanning 24 years, a logistic regression model was fit with proportion promoted as the dependent variable and gender, race, age, education, and grade as independent, or predictor, variables. Individual models were fit to distinct categories of predictor variables, emphasizing gender and race effects, and panel plots of observed and modeled proportions, along with model parameter estimates, were provided to principle investigators and researchers for review and use in further research into U.S. federal workforce equality. Data access, management, and retrieval were executed in SQL; statistical programming and graph development were executed in R, specifically with glm and ggplot.

Development of a quantile regression model to estimate the effect of gender on wages within the U.S. federal workforce. Various studies have been conducted to estimate an apparent difference in wages between female and male federal workers. This model was used in a study that showed a reduction in gender wage differential, as compared to other studies, when age, race, education, agency, and occupation are controlled for. Regression models were fit to approximately 50,000,000 historical federal employee career and pay records, spanning 24 years, to estimate the 10^{th}, 50^{th} (median), and 90^{th} percentiles of employee pay given various combinations of age, race, education, agency, and occupation. Agency and occupation were treated as fixed effects, with thousands of distinct levels, causing a computationally intensive problem involving designs of dimension up to 50,000,000 X 2,500, and solution was found to be feasible and efficient only through use of column-wise indicator-sum methods (very efficient with low density factors), use of sparse matrices, and programmatic matrix construction (as opposed to the standard R method of as.matrix.csr). Panel plots were produced to compare models and gender effect estimates were compared to published values from external research. A future project will model parameter estimate probability distributions and measure statistical significance through simulation, using estimates from a series of models fit to random samples of observations. Data access, management, and retrieval were executed in SQL; statistical programming and graph development were executed in R, specifically with SparseM, rq.fit.sfn, and ggplot.

Development of an upper bound control limit for multi-sensor air flow discrepancies in selective catalytic reduction (SCR) tests. SCR reactor tests require up to 48 hours to complete, throughout which, airflow in the reactor chamber must be controlled to regulate catalytic exposure to various gas mixtures. Airflow is sampled by two primary sensor systems and if a significant discrepancy is observed, investigation and intervention are prescribed. Data were presented by the responsible system engineer after unsuccessful attempts at modeling confidence bounds, using standard polynomial regression methods available in the corporate statistical program.

Key methods and results:

- The sensor groups have a known mechanical relationship, such that the results from one group are always greater than the other – it is the ratio of this discrepancy that is monitored
- A simple X-Y plot (Y=ratio, X=lower sensor group) revealed a readily recognizable exponential form
- Analysis of the upper bound of Y and the lower bounds of X and Y indicated shift and scale requirements on both axes, with a resulting model of $\varepsilon + \alpha\mathbf{e}^{\beta(x+\delta)} = \varepsilon + \alpha\mathbf{e}^{\beta\delta}\mathbf{e}^{\beta x} = \varepsilon + \alpha'\mathbf{e}^{\beta x}$
- Logarithmic transformation was attempted in hopes of converting the problem to single parameter linear regression, but simplification was hindered by ε remaining in log Y values
- Normal sum of squared differences equations were derived, one for each parameter, along with their first derivatives

- The three derivative equations were iteratively solved for zero using Newton's method of roots, starting values taken from the boundary analysis, and the second derivatives of the normal equations, until values for ε, α', and β were identified that provided simultaneous optimal solutions
- To simulate ε, α', and β variation, 1,000 solution sets were generated based on randomly selected observation pairs
- To simulate Y variance, 5,000,000 X,Y values were simulated by randomly selecting ε, α', and β sets and applying them to all recorded X levels in 10,000 iterations
- The upper .95 confidence limit was taken for each X value (the least value of the upper 5%)
- Final ε, α', β values were derived using the partitioned .95 boundary – the exponential model with these parameters is the new upper control limit
- Upper control limit suitability was confirmed by measuring the proportion of actual observations above it
- The new control limit was placed in service
- Programming was accomplished in R and SQL

Customer service, dispatch, and cost management. On-time delivery, honoring customer commitments, and reducing freight costs are significant operational objectives. Customer retention is an important revenue factor while cost of delivery is 20% or more of invoiced revenue. A team was formed to study delivery assignment with the objective of improving freight utilization and on-time delivery while reducing overall freight costs.

Key features and methods:

- Deliveries are scheduled by customer service representatives, while the customer is on the phone, within the enterprise resource planning (ERP) system
- To reduce the overall number of deliveries required, new SQL programs were developed to scan scheduled loads, at time of scheduling, for potential delivery combination based on product weight, truck capacities, and vicinity of destination
- Each combination eliminates a distinct delivery, reduces immediate operating costs, and improves long range pricing competitiveness
- Additional immediate information-action triggers attempt to move inventory and eliminate future deliveries by scanning customer orders for in-stock material that has not yet been requested but could be combined with scheduled or customer-inquiry deliveries
- These answers to operational improvement questions are "entirely free," using information that results from daily customer service activities

Risk reduction in a stochastic decision process setting. Developed a Markov decision process to model the economic impact of decision to launch or secure vessels with threat of approaching storm (fishing industry, submitted for review). The model improves risk analysis and contributes to economic and safety gains through use of appropriate probability models derived from past data .

Key features and methods:

- Storm stage, strength, and vicinity transition probabilities result from Markov chain derived from historical hurricane stage, category, and vicinity transition data
- Damage estimates are from regression of historical damage on actual storm category and vicinity
- Probabilistic/deterministic launch decision models are specifiable as study parameters for risk analysis
- Optimal policies are identified through iterative storm development/policy application assessment
- Programmed in SAS (procs SQL, REG, IML, and GPLOT for data aggregation, regression, matrix operations, and graphs; macro development to bind procs into single-call commands)

thomas.balmat@duke.edu **Thomas J. Balmat** 919-820-1057

Reduction of variance in a nonlinear, non-normal process setting. Developed variance models for known non-linear process functions in several variables, some of which exhibit non-normal probability distribution.

Key features and methods:

- An approximating Taylor polynomial is derived and programmed using partial derivatives of the study system physical model
- System variance is estimated from approximating Taylor polynomial using propagation of moments
- The variance estimating function is optimized (mean levels for independent parameters are specified that achieve minimal expected system response variance) using nonlinear programming methods
- Independent parameter probability distributions are derived at system-variance optimal levels by matching moments or percentiles to observed data (distributions include normal, lognormal, exponential, parabolic, and gamma)
- System response is simulated using independent parameter optimal levels and derived probability distributions (Monte Carlo simulation)
- A non-normal probability model is fit to the simulated response data by matching moments (up to four) and analyzing probability plots (fitted distributions include normal, lognormal, exponential, and Johnson S family)
- Operating confidence intervals are developed using numerical percentile estimates for the chosen distribution given parameters from the moment matching step
- Programmed in SAS (data steps, various random number functions, cumulative mass, inverse mass, math functions, macros, and procs SQL, NLP, IML, and GPLOT)

Custom plant management information system to coordinate all critical activities from initial bid to invoicing. With this system customer, project, and product information is no longer “lost” and product is delivered on time, as ordered. Automated business transaction monitoring and management guides daily activities toward completion of critical events using customer promise dates, project updates, plant capacity, and product requirements imposed by customer, municipal, and departments of transportation specifications.

Inventory location assignment. Finished items are automatically assigned a storage location as they are scheduled using known location capacities and projected inventory activity (receipts from production schedule, shipments from shipment schedule). Location is printed on identification label affixed to product and is listed on dispatch loading instructions. Material is no longer “lost” or “hunted for.” With no additional human effort, the system instructs where to place and find an item.

Inactive inventory reassignment. Occasionally, due to changes in a customer project, serviceable material is not delivered. This module identifies such items and searches the daily production schedule for possible application, eliminating the need for remanufacture. Comparison is based on material properties, physical geometry, and custom features. No additional human activity or information is required for this; inventory, bill of material, and schedule records from daily operations along with material property models provide “free” answers to an important material efficiency question.

Dispatch management. Freight costs are factored into product pricing and often equal 20% or more of invoice revenue. To reduce overall number of deliveries required, scheduled loads are scanned for potential delivery combination based on product weight, truck capacities, and vicinity of destination. Each combination reduces required deliveries by one and improves the competitiveness of our pricing while reducing operating costs. Companion reports assist customer service agents with delivery commitment while the customer is on the phone, again scanning for combination to control costs. One variation proactively moves inventory and eliminates future loads by scanning customer orders for in-stock material that has not yet been requested but could be combined with scheduled or customer-inquiry deliveries. These answers are also “free,” using information that results from daily sales, manufacturing, and delivery activities.

Employee incentive pay. Production employees are paid an incentive bonus for hours worked in departments that achieve their target level of productivity. Employee hours and departments worked in, recorded with electronic time clocks, are combined with inventory receipt data; bonus levels and pay rates are calculated by

employee then appended to the weekly payroll submission. This is another “free” answer to an important efficiency question, since all required information is taken from daily time and attendance and inventory activities. An important variation of this module is a daily version that predicts, as the week advances, eventual bonus levels. Crew leaders communicate this to employees, thus proactively achieving employee satisfaction and required efficiency.

Customer telephone call and website inquiry management. Incoming telephone calls and web site inquiries trigger a query to the customer contact database. When the caller ID or web supplied identifying information is recognized, calls and messages are forwarded to the service representative assigned to that customer. Known contacts don’t get “lost” in the phone system and inquiries are not “dropped.” Current customer order and delivery activity are displayed during conversation.

Company web site. To expand our market and improve customer information and relations we developed a web site, hosted in-house, using graphics and overall design from outside and catalog, product, and service information, HTML development, and programming from within. SQL Server database interfaces retrieve and present select internal information to the world. See www.carolinaprecast.com

Oracle E-Business integration. In 2007, the company was purchased by Hanson Pipe and Precast. At that time all invoicing and inventory valuation functions were transferred to the corporate office using Oracle EBusiness Suite. Sales orders, delivery bills of lading, and product bills of material are maintained locally by plants and stored in a centralized Oracle database. Significant Oracle projects include:

- Develop custom SQL procedures to export sales order and bill of material records from local plant management SQL Server databases to the corporate Oracle database.
- Develop monitoring and update SQL procedures to synchronize customer credit status, sales line value, shipment and invoice value, and bill of material configuration between local plant management systems and the corporate Oracle database.
- Maintain, report, and improve accuracy of bills of material, inventory transactions, and inventory valuation for 20 pipe and precast plants throughout the eastern United States.
- Development platforms and systems interfaced to: Microsoft SQL Server, SQL Server T-SQL, Microsoft
- Access, Microsoft Visual Basic, Microsoft Visual Studio, Oracle SQL, Oracle E-Business Suite, SAP Business Objects, Adobe Dreamweaver, Adobe Photoshop, Adobe Flash, HTML, Java, Active Server Pages (ASP), Microsoft Internet Information Server, Infotronics Attendance Enterprise (time and attendance), Vertical Systems Televantage (Windows based telephone pbx), SBT Accounting, PayChex Payroll.

Model fitting and robustness testing. Visually graph historical process data and select appropriate probability distributions, mainly normal and Johnson S-curves. Calculate best-fit distribution parameters using sample mean, variance, skewness, and matching of higher order moments. Assess goodness of fit with probability plots and outlier analysis. Specify probable process limits from distribution confidence limits calculated from distribution cumulative mass function. Present findings in formal reports for customer use.

Design of experiments. With assistance from engineers, identify critical process responses and controllable factors and ranges. Develop minimal run factorial experiments (fractional, blocked, paired, etc.) and assist with execution and data collection. Identify significant factors, interactions, and response models using regression, ANOVA, factor subset analysis, PRESS, etc. Challenge models and study outliers and influential observations using normality plots, Cook’s D, etc. Explain models and assist with implementation and process improvement.

Sample plans. Specify minimal quantity inspection sample sizes using historical process variation and binomial probabilities (OC curves).

Quality information system. Develop programs and queries for data retrieval, organization, analysis, and reporting. Numerical programs for matrix operations, cumulative probability functions, linear regression, combinatoric functions, and non-parametric methods. Frequency distribution plots, response surface graphs, and observed/model probability plots for visual analysis. Platforms: BBN RS/1, MS Access, Visual Basic, Oracle, SQL, Excel.