Air Force Institute of Technology

AFIT Scholar

Faculty Publications

Forecasting US Army Enlistment Contract Production in Complex Geographical Marketing Areas

Joshua L. McDonald
Army Materiel Systems Analysis Activity

Edward D. White
Air Force Institute of Technology

Raymond R. Hill
Air Force Institute of Technology

Christian Pardo
Defense Threat Reduction Agency

Follow this and additional works at: https://scholar.afit.edu/facpub

Part of the Business Analytics Commons, and the Operational Research Commons

Recommended Citation

McDonald, J. L., White, E. D., Hill, R. R., & Pardo, C. (2017). Forecasting US Army enlistment contract production in complex geographical marketing areas. Journal of Defense Analytics and Logistics, 1(1), 69–87. https://doi.org/10.1108/JDAL-03-2017-0001

Forecasting US Army enlistment contract production in complex geographical marketing areas

Joshua L. McDonald

Army Materiel Systems Analysis Activity, Aberdeen Proving Ground, Maryland, USA

Edward D. White

Department of Mathematics and Statistics, Air Force Institute of Technology, Wright Patterson AFB, Ohio, USA

Raymond R. Hill

Department of Operational Sciences, Air Force Institute of Technology, Wright Patterson AFB, Ohio, USA, and

Christian Pardo

Defense Threat Reduction Agency, Kirtland Air Force Base, New Mexico, USA

Abstract

Purpose – The purpose of this paper is to demonstrate an improved method for forecasting the US Army recruiting.

Design/methodology/approach – Time series methods, regression modeling, principle components and marketing research are included in this paper.

Findings – This paper found the unique ability of multiple statistical methods applied to a forecasting context to consider the effects of inputs that are controlled to some degree by a decision maker.

Research limitations/implications – This work will successfully inform the US Army recruiting leadership on how this improved methodology will improve their recruitment process.

Practical implications – Improved US Army analytical technique for forecasting recruiting goals..

Originality/value – This work culls data from open sources, using a zip-code-based classification method to develop more comprehensive forecasting methods with which US Army recruiting leaders can better establish recruiting goals.

Keywords Regression, Comparative studies, Sales forecasting

Paper type Research paper

Introduction

Since the formal elimination of the draft by Congress in 1973, the USA Army has maintained an All-Volunteer Force (AVF) (Waddell, 2005, pp. 413-442). At the lowest echelon of the Army recruiting system, US Army Recruiters are tasked to help fill the ranks of the AVF by actively pursuing qualified future soldiers with the ultimate goal of generating required enlistments. At higher echelons of recruiting leadership, however, a fundamental concern arises in how best to distribute recruiting goals to subordinate recruiting echelons throughout the country.

Implied in the latter concern is an assumption that a quantitative relationship between numerous recruiting factors – both within and outside of the recruiters' control – and enlistment production in each geographical recruiting area exits, can be satisfactorily established and can be exploited for forecasting purposes. If so, recruiting leadership can then maximize potential recruiting contracts by altering recruiting goals subject to a set of organizational constraints. Thus, the two primary steps of setting recruiting goals consist of:

- defining an appropriate quantitative relationship that is robust to extrapolation into the future; and

- maximizing the outputs based on organizational constraints and the quantitative model parameters found in the first step. We use the acronym US Army Recruiting Command (USAREC) to refer to the corporate, goal-setting body of Army recruiting leadership.

This article focuses on the first of recruiting leadership's tasks – the development of a quantitative relationship between recruiting market factors and enlistment production – as it is fundamental to successful execution of the second task, optimizing production. We are not the first to do this; USAREC has used and currently uses such quantitative models in the past. We do however offer an improved methodology for achieving a useful quantitative model.

We first establish the extent to which we can accurately forecast the relationship between enlistment supply and demand factors and enlistment contract production. We assume as supply factors those which are outside of the recruiters' control; local unemployment rates are an example. Enlistment demand factors, by contrast, are those over which the Army and recruiters have some control; numbers of recruiters on-hand and recruiting goals are examples. In terms of the response, we focus specifically on Regular Army (RA) enlistment contracts in 38 recruiting regions that span the USA, territories excluded. We subdivide the enlistment contracts into three mutually exclusive categories of interest to USAREC:

(1) high-aptitude high school graduates (GA);

(2) high-aptitude high school seniors (SA); and

(3) all others (OTH) (Flesichmann and Nelson, 2014).

These are not the standard responses used by USAREC in their forecasting model, but we address in this paper why changing to these make sense.

We analyze data from both recruiting organizations and open sources for the period Fiscal Year (FY) 2010-2014. We take advantage of open source data at the county level and map this county-level data to each ZIP code-based recruiting market boundary. To complete this mapping, we introduce a way to compress data from over 3,000 counties and 42,000 ZIP Codes into 38 markets. We then apply principal components analysis (PCA) and mixed

stepwise regression to the re-mapped data to develop adequate, parsimonious models for each recruiting market and contract type. The application of PCA represents a significant contribution to the level of statistical rigor in our model development methodology over previous efforts.

Quantitative prediction models are useful when they yield accurate predictions. We use hold-out samples for model validation. We use only the first 75 per cent of the data to estimate ordinary least squares models; the latter 25 per cent of the data are used in forecast validation. To obtain a better appraisal of model stability during validation, we create additional realism by using simple linear trend forecasts of market supply variables. At the conclusion of this step, we achieve our penultimate objective by rendering quantitative market- and contract-specific comparisons of model performance within the context of a forecasting scenario. As a brief concluding excursion, we compare our multivariate forecasting approach to several univariate time series models.

In summary, we introduce an improved methodology for model development and assessment. To this end, we introduce an improved way to generate appropriate response data, and we demonstrate the use of principle components analysis to reduce model dimensionality and demonstrate an improved level of statistical rigor in our model development methodology. We also present an empirical study comparing our model to other common models.

Background
There is a vast literature on military recruitment, enlistment and retention. We focus on macro- and micro-economic studies. Macroeconomic studies are highly aggregated geographically, involving only a handful of national regions. Microeconomic studies are geographically disaggregated, typically at the ZIP code level. Due to space limitations, we here provide general findings based on a detailed discussion found in McDonald (2016).

Our review of pertinent literature on Armed Forces recruiting spans a 26-year period from 1985 through 2011. In total, macroeconomic studies of enlistment supply are helpful in describing the "big picture" of recruiting models. There seems to be some general agreement between these studies in the significance of a few select factors: unemployment, qualified military available population, veteran population and recruiter strength are of particular importance in this regard (Asch et al., 2009; Warner et al., 2001; Dertouzos and Garber, 2006, 2008; Dertouzos, 1985). However, these works have limited capacity to predict contract production for geographical areas corresponding to recruiting market boundaries. We therefore acknowledge the added value of microeconomic models in both their geographic specificity and use of validation data despite their reduced specificity in the response and difficulties in comparing fit adequacy (Gibson et al., 2011, 2009).

There are useful conclusions derived from the prior literature. First, the literature is dominated by econometric methodology. The econometric studies share a common objective of describing socioeconomic effects on recruiting over time. In nearly every study, this involved some form of regression. We use a similar quantitative modeling methodology albeit with useful extensions.

Second, there is some broad agreement that several factors are correlated with recruit production. Of these, geography appears to be statistically significant (Dertouzos and Garber, 2006, 2008). Unfortunately, we cannot conclude how non-geographic factor effects change with respect to geographic location; each study uses different geographic boundary definitions, none of which correspond to actual USAREC definitions. In any case, geography is statistically significant based on the total body of empirical results. We do note with some caution that statistics gathered at the microeconomic level may have greater measurement

error (Murray and McDonald, 1999). Thus, our methodology seeks an appropriate balance between required geographic specificity with respect to recruiting markets and data measurement errors inherent in higher resolution. We maintain a variable set broadly consistent with previous literature to provide a general basis for comparison.

Interestingly, relatively little of the previous research specifies predictive models designed to produce forecasts into future time periods. Recruiting – like any private-sector marketing effort—requires decision-making (i.e. an irrevocable allocation of resources) in the face of uncertainty (Howard and Abbas, 2015, pp. 1-19). While the studies we reviewed provided some indication of how variables respond to time, most did not explicitly describe model performance in an out-of-sample time period or provide any kind of probabilistic statement regarding future behavior. This observation is a primary motivator for our methodology, specifically with regard to our model validation efforts.

Methodology
Data description and cleaning
We collected data from USAREC and open sources. Our goals were to use variables mentioned in previous literature and provide a representative sample of pertinent supply and demand factors. In all, we collected data on 26 separate metrics for the period FY10 to FY14. See Table I for the definition of these first 26 variables. While data from recruiting leadership was available at the market level, open source data were collected almost universally at the county level. This presented us with a fundamental difficulty because recruiting market boundaries – defined by a set of ZIP codes – are incompatible with political boundaries.

Because the market level data provided the greatest level of resolution, we devised a method of weighting county-level data to express it in terms of recruiting market boundaries. Let $Z_i \subseteq Z$ be the subset of ($m = 1, 2, 32846$) ZIP Code Tabulation Areas (ZCTAs) within each unit i boundary. Let C be the set of ($n = 1, 2, 3141$) counties in the USA. Let C_i' be the set of counties which intersect with a ZCTA in a unit's area of responsibility $\left(C_i' \subseteq \{C \cap Z_i\} \right)$. We then define a weighted statistic $z^{'}$, for unit i (scripts j and t are omitted because this definition applies to all variables and times) as:

$$z_i{}' = \sum_{m \in Z_i} \sum_{n \in C_i'} \upsilon_{m(n)} z_n \tag{1}$$

where I $\upsilon_{m(n)} \equiv$ the proportion of county n population residing in ZCTA m, from the 2010 Census and $Z_n \equiv$ the available discrete statistic for county n, where $|z_n| \geq 1$ (USA Census Bureau, 2015).

We implemented equation (1) only for fractional data, applying [equation (1)] separately to the numerator and the denominator prior to dividing. We explored weighting a raw value such as population but found that aggregating to market levels produced a total value greater than the original. This is likely due to some double-counting in our formulation of z_i'. However, we assume that similar over-estimation errors applied to the numerator and denominator of a single rate are likely to cancel each other out. The reasonability of our resulting weighted values further increased our confidence in this method.

We also took care to ensure our sample size of time series data were adequate. A common rule of thumb is to have at least 50 observations for model estimation (Montgomery et al., 2015, p. 39). Recruiting data in the model estimation set provided a suitable $N = 45$ observation. However, much of the county-level data proved available only at annual,

Index	Variable name	Description
1	Voter participation rate	Votes cast for president/total adult population (2008 and 2012, County)
2	Sponsor share	Number of Army active duty sponsors/total active duty military sponsors (2010-2013, Annual, ZIP code)
3	Labor participation rate	Persons in labor force/total working-age population (2010-2014, Annual, County)
4	Unemployment rate	Employed persons/persons in labor force (2010-2014, Monthly, County)
5	Cohort HS graduation rate	Graduates from freshman high school class/size of freshman class (2010-2014, Annual, County)
6	Violent crimes	Number of violent crimes (2010-2014, Annual, County)
7	Obesity	Number of obese persons/total population (2010-2014, Annual, County)
8	Illicit drug use	Number of persons using illicit drugs/total Population (2010 and 2012, County)
9	Urban population rate	Number of persons in urban zones/total population (2006 and 2013, County)
10	Propensity	Number of youth inclined toward military service (2010-2014, Semiannual, Battalion)
11	QMA population	Number of youth aged 17-24, qualified without a waiver (2010-2014, Annual, ZIP Code)
12	17-24 Population	Number of youth aged 17-24 (2010-2014, Annual, ZIP Code)
13	Battalion recruiting station identifier (RSID)	Recruiting battalion boundaries (2010-2014, Annual, ZIP Code)
14	Lag-1	Number of total contracts produced from previous month (2010-2014, Monthly, Battalion)
15	Reg. Army GA Mission	Goal for number of GA contracts (2010-2014, Monthly, Battalion)
16	Reg. Army SA Mission	Goal for number of SA contracts (2010-2014, Monthly, Battalion)
17	Reg. Army OTH Mission	Goal for number of OTH contracts (2010-2014, Monthly, Battalion)
18	Reg. Army GA Achieved	Number of adjusted GA contracts produced (2010-2014, Monthly, Battalion)
19	Reg Army SA Achieved	Number of adjusted SA contracts produced (2010-2014, Monthly, Battalion)
20	Reg. Army OTH Achieved	Number of adjusted OTH contracts produced (2010-2014, Monthly, Battalion)
21	Contract share	Number of Army contracts/all DoD contracts (2010-2014, Monthly, Battalion)
22	Recruiter share	Number of Army recruiters/all DoD recruiters (2010-2014, Monthly, Battalion)
23	Army recruiters	Number of Army active and reserve recruiters based on PERSTAT (2010-2014, Monthly, Battalion)
24	Appointments made	Number of appointments scheduled and reported to USAREC (2010-2014, Monthly, Battalion)
25	Appointments conducted	Number of appointments conducted and reported to USAREC (2010-2014, Monthly, Battalion)
26	Processing days	Number of days to process recruits (2010-2014, Monthly, Battalion)

Table I.
Variable names and definition of original 26 variables considered

semi-annual or quarterly intervals. We therefore expanded the county-level data to that of its recruiting counterpart by applying stochastic mean value imputation to the county data. This involved creating random realizations of monthly county-level data points along a trend-line between the observed annual data (Montgomery *et al.*, 2015, pp. 18-19).

To illustrate the imputation, let z'_t and z'_{t+12} be realizations of a market-level statistic at the same month t in subsequent years; the in-between monthly values of $z'_{t+2}, z'_{t+3}, \ldots, z'_{t+11}$ must be imputed. We obtain the mean values for all imputed t, $\hat{\mu}_t$, by subtracting z'_t from z'_{t+12} and dividing by 12 to obtain the gradient, δ such that $\hat{\mu}_t = z'_t + \delta t$ for $t = 1, 2, 12$.

A conservative standard deviation $\hat{\sigma}$ is then $\hat{\sigma} = (12\delta)/4 = 3\delta$ (Wackerly *et al.*, 2008, pp. 10-11). The parameters $\hat{\mu}_t$ and $\hat{\sigma}$ are then used to impute the random realizations for each $z'_{t+1}, z'_{t+2}, \ldots, z'_{t+11}$ by sampling from a normal distribution.

Modeling methodology
A common approach in the past research builds a linear regression or time series model involving numerous potential variables, seemingly without regard to multicollinearity. We used a more rigorous modeling process.

Variables 17-19 in Table I are used as dependent variables with all other variables as potential predictor or independent variables. Principle component analysis provided the basis for the response choice. Principle component analysis is also used to reduce the dimensionality of the predictor set of variables (and thus reducing multicollinearity when using the full set). The reduced set is modeled based on stepwise regression techniques using R^2_{adj} to judge model fit and studentized residuals to conduct residual analysis when assessing model adequacy. Model adequacy involved residual plots to assess constancy of variance and reasonable normality, whereas measures such as Cook's distance and Hat-matrix elements were used to examine potential outliers or influential points. Details of these analyses are found in McDonald (2016).

Validation of forecasts

Data splitting
The primary purpose of our models is to predict future data. Model validation examines how well the estimated model performs in the presence of future data. Data splitting is used to conduct model validation. Observations $t = 1, 2, T$ define the estimation set used in the model building processes, whereas observations $t = T + 1, T + 2, T + \tau$ define the validation set. In our data set, we let $T = 45$ and $\tau = 15$ as 15 to 20 observations are recommended to gain an adequate assessment of prediction performance (Montgomery *et al.*, 2012, p. 375) and recruiting leadership begins setting missions a few months prior to the next full recruiting year. By adding three months to the validation set, we effectively re-create the decision situation from the leadership's point of view, predicting contract production over an extended planning horizon using only the data realized by the forecast origin, T.

Forecast metrics
The usual metrics of model fit such as R^2_{Adj} and R^2_{Pred} do not apply as forecast accuracy metrics. The following two metrics are used for assessing forecasting accuracy:

$$\text{MAD} = \frac{1}{N} \sum_{t=T+1}^{T+\tau} \left| y_t - \hat{y}_t \right|$$

$$\text{MAPE} = 100\% \cdot \frac{1}{N} \sum_{t=T+1}^{T+\tau} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (2)$$

where y_t is the actual response at lead time $T + 1, T + 2, T + \tau$ and $\hat{y}_t$ is the predicted value of the same (Montgomery *et al.*, 2015, pp. 64-74). Because the mean absolute deviation is scale-dependent, we also use the mean absolute per cent error.

We assume independent model and variable forecast errors for a one period-ahead forecast, providing a 100 per cent $(1 - \alpha)$ prediction interval as:

$$100(1 - \alpha)\% \text{ PI} = \hat{y}_{t+1} \pm z_{\alpha/2} \sqrt{\hat{\sigma}_a^2 + \hat{\beta}^2 \hat{\sigma}_x^2} \tag{3}$$

where $z_{\alpha/2}$ is the $(1 - \alpha)$ per cent critical value, $\hat{\sigma}_a^2$ is the variance estimator regressing e_{t-1} on e_t, $\hat{\beta}^2$ is the square of the coefficient for x obtained from the model and $\hat{\sigma}_x^2$ is the NID(0, σ^2) estimate of error variance obtained from the first-order autoregressive model of x (Montgomery *et al.*, 2015, pp. 202-204). We generalize the results found in equation (3) for multiple independent regressors to obtain:

$$100(1 - \alpha)\% \text{PI} = \hat{y}_{t+1} \pm z_{\alpha/2} \sqrt{\hat{\sigma}_a^2 + \sum_{x_j \in \hat{\mathbf{x}}} \hat{\beta}_j^2 \hat{\sigma}_{x_j}^2} \tag{4}$$

where we denote the set of forecast regressors with $\hat{\mathbf{x}}$.

To implement equation (4) in a predictive role, simple trend models are used to extrapolate independent factors into the future prediction periods.

Empirical results

Response and variable determination

The first phase of the work examined the current USAREC modeling approach. The current approach involves a response constructed using some of the data in Table I. Using the variable index indicated in Table I the output response used is:

$$y_1 = \frac{x_{18} + x_{19}}{x_{23}} = f(x_4, x_{10}, x_{13}, x_{27}, x_{28}, x_{29}) \tag{5}$$

where:

$$x_{27} = (x_{15} + x_{16} + x_{17})/x_{23} \tag{6}$$

$$x_{28} = (x_{15} + x_{16})/x_{23} \tag{7}$$

$$x_{29} = (x_{18} + x_{19} + x_{20})/x_{23}. \tag{8}$$

The current USAREC model is relatively parsimonious and produces a good fit ($R_{Adj}^2 = 0.93$). The model including x_{29} provides a somewhat better fit. However, it contains a rather concerning issue in its formulation: the predictor x_{29} is almost identical to the response, y_1. To improve upon the USAREC approach, we therefore reconsidered how to approach the entire modeling process.

The first step was to reconsider the range of responses available. There are seven potential responses as summarized in Table II. Note the first response comes from the current USAREC approach and the second response is the variable x_{29} removed from the current model. The other five are responses are found in Table I.

A principle components analysis was conducted on these responses and revealed loadings on two quantities, one being y_4. Because responses y_1, y_2 and y_7 are all ratio values, all involving the x_{19} variable, the logical choice was to use responses y_3, y_4 and y_5. These provide meaningful responses for the model development and are easy to understand. We must note that the independent formulation of SA achieve, y_4, as a response variable constitutes a remarkable departure from the current USAREC model, which lumps SA and GA Achieved (y_3 and y_4) together in the same response. Further discussion is devoted to why this is a sensible departure. Nevertheless, subsequent model development focused on each of these three responses, GA, SA and OTH production (i.e. contracts achieved), respectively.

Multiple regression methods using the three responses and the remaining 23 variables from Table I resulted in serious problems with multicollinearity. A principle components analysis was again used this time to reduce the dimensionality of the independent variable set. As Figure 1 indicates, five components account for approximately 79 per cent of the variance. The initial set of loadings are provided in Table III. Using the loadings and factors, along with context knowledge of the problem and a few iterations of principle component analysis, we finalized the set of five component variables as x_4 along with the derived variables:

Table II.
Responses considered in re-examination of a casual model for recruitment forecasting

Responses	How defined using Table I variables
y_1 (GSA_PR)	$(x_{18} + x_{19})/x_{23}$
y_2 (Vol_PR)	$(x_{18} + x_{19} + x_{20})/x_{23}$
y_3 (GA Achieved)	x_{18}
y_4 (SA Achieved)	x_{19}
y_5 (OTH Achieved)	x_{20}
y_6 (GA + SA Achieved)	$x_{18} + x_{19}$
y_7 (Contract Share)	x_{21}

Note: Labels for each of the variables used found in Table I

Figure 1.
Principal component eigenvalues and Horn's curve for independent variable candidates
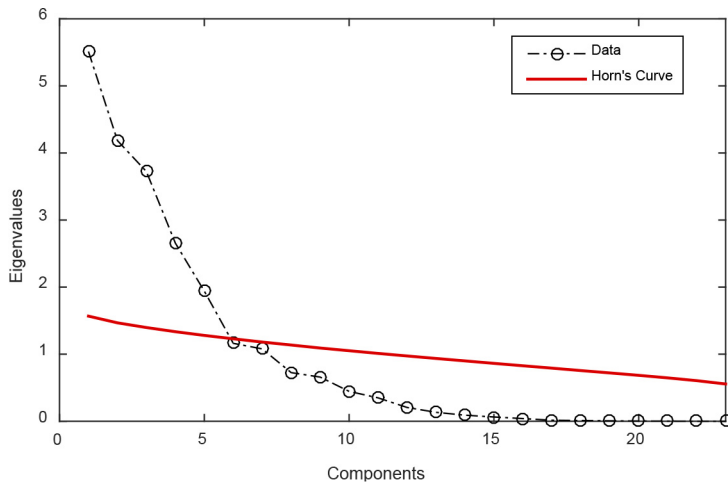
Variable	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
x_1	0.4724	*0.6486*	−0.0628	−0.0948	0.2746
x_2	0.5045	*0.7153*	0.0180	0.0579	−0.0407
x_3	*−0.7619*	0.2076	0.2354	−0.3092	0.2346
x_4	−0.0377	−0.4225	*−0.6136*	−0.0599	0.5270
x_5	−0.4288	*0.5901*	0.4573	0.0041	−0.1782
x_6	−0.0709	−0.2936	*−0.7065*	−0.3887	0.0780
x_7	0.5014	*0.7327*	0.0519	0.1881	−0.0123
x_8	*−0.7177*	−0.4122	0.0862	0.1361	0.2085
x_9	−0.4932	*−0.7896*	0.0706	−0.0925	−0.2529
x_{10}	*0.6309*	−0.4604	−0.2794	0.2299	−0.3635
x_{11}	*−0.6595*	0.3620	0.2762	−0.4695	0.1842
x_{12}	*−0.7745*	0.1379	0.3073	−0.4403	0.0694
x_{15}	*0.5548*	−0.2439	0.2772	−0.4699	0.4496
x_{16}	0.2977	−0.1260	*0.7181*	0.1922	0.1013
x_{17}	*0.7237*	−0.2960	−0.0005	−0.2013	0.2052
x_{22}	0.3273	0.1584	−0.2562	*−0.6416*	−0.0017
x_{23}	0.3109	0.2602	−0.5920	−0.5654	0.0165
x_{24}	0.2679	−0.1399	0.3664	*−0.7385*	−0.3609
x_{25}	0.3141	−0.2481	0.3994	−0.5156	*−0.5592*
x_{26}	−0.0334	0.0158	0.0654	*0.1292*	−0.0716
x_{27}	0.5856	−0.3851	*0.6286*	−0.0026	0.3038
x_{28}	0.3848	−0.2956	*0.7824*	0.0378	0.2950

Notes: Labels for variables found in Table I; Italics indicate the larger loadings

Table III.
Principle components analysis summary for initial variable set

$$x_{30} = x_{25}/x_{24} \tag{9}$$

$$x_{31} = x_{15} + x_{17} \tag{10}$$

$$x_{32} = x_{10}/x_{23} \tag{11}$$

$$x_{33} = x_{12}x_{10}x_5(1 - x_7) \tag{12}$$

In general terms, x_4 captures unemployment rates, x_{30} the ratio of appointments (a surrogate for recruiter effort), x_{31} the combined GA and OTH mission, x_{32} the SA contracts per recruiter and x_{33} a measure of the young adults available. This reduction from 23 potential variables to just 5 variables is an important modeling consideration.

To conclude this portion, we provide the final correlation matrix for the five variables selected. Of the ten off-diagonal elements in Table IV, seven are less than $|0.2|$. We note

Variable	x_4	x_{30}	x_{31}	x_{32}	x_{33}
x_4	1	−0.213	0.196	−0.370	−0.086
x_{30}		1	−0.130	0.199	0.409
x_{31}			1	0.073	0.079
x_{32}		*symm*		1	0.077
x_{33}					1

Table IV.
Correlation matrix R for the reduced set of independent variables
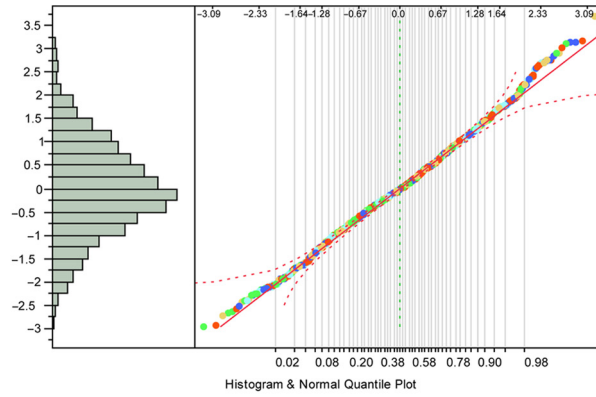
(a)
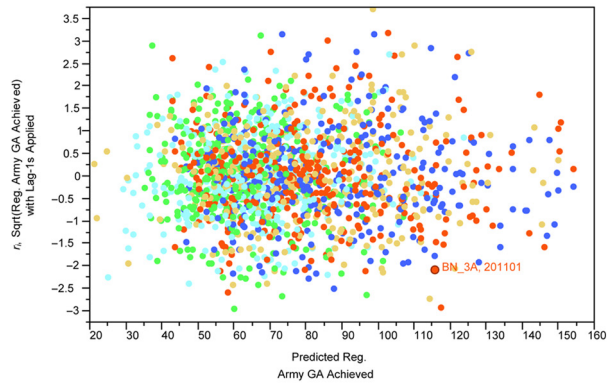
(b)
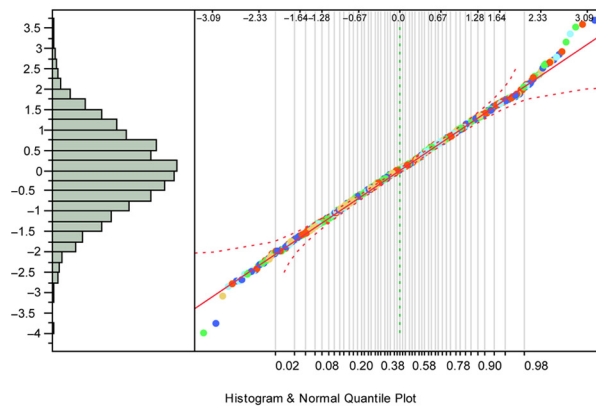
(c)

Figure 2.
Left side are quantile plots showing general acceptance of normality assumption. Right side are plots indicating no issues with the constant variance assumption

(*continued*)
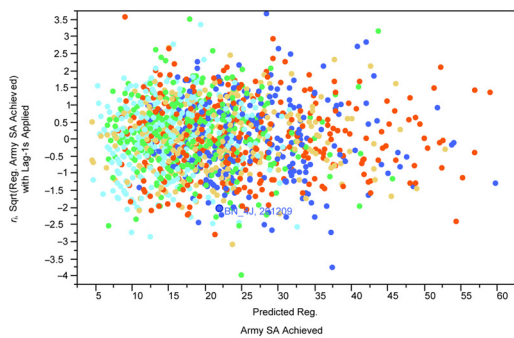
(d)
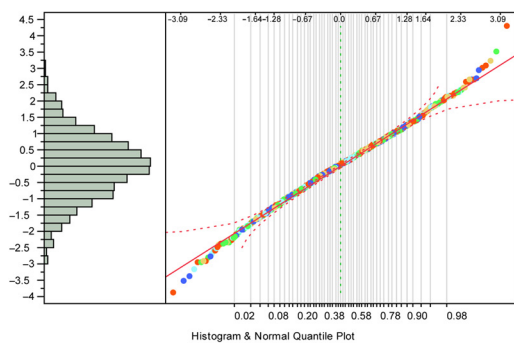
(e)
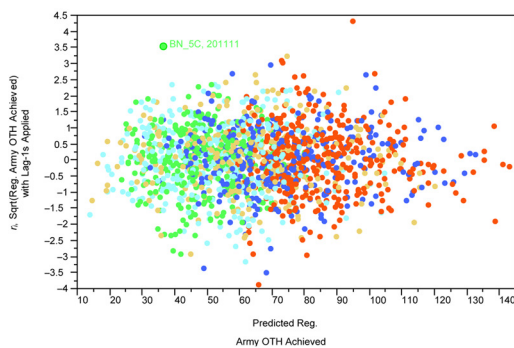
(f)

Notes: These are for the GA, SA and OTH responses
from top to bottom; subfigures (a), (c) and (e) are
histograms and normal probability plots of Army GA,
SA, and OTH Achieved, respectively. Subfigures (b),
(d) and (f) are residual versus predicted plots for Army
GA, SA, and OTH Achieved, respectively

Figure 2.

entries of 0.37 and 0.41, but these are still both less than 0.5 and are deemed not overly troublesome. Thus, we are confident that this reduced set of five variables are adequate for final model development.

Mixed stepwise selection

An initial examination via stepwise regression ruled out quadratic models or linear models with interaction terms. Thus, a pure first-order model is used. Further, based on Box-Cox analyses, all responses were transformed via the square root transformation to improve compliance with the constant variance assumption of the residuals. Despite the transformation, autocorrelation remained present in the residuals. Adding a first-order autoregressive term to the model alleviated this concern. Finally, we needed to model each of the battalions within USAREC. This was done using indicator variables for each market, while simultaneously allowing for the predictor effects to change between markets (i.e. we modeled categorical-continuous interactions). The final models, while seemingly complex, facilitated model adequacy analyses.

Figure 2 provides the pertinent results of the model adequacy analysis. The models for GA, SA and OTH contracts are ordered top to bottom. On the left, we see no reason to doubt reasonable normalcy of the residuals and on the right side we see no reason to doubt the constancy of variance. Residual analysis for outliers and influential points revealed nothing of concern. Table V provides the summary of each model fit. While value of p, the number of terms in the model, is high, the vast majority of these are the indicator variables used to derive the individual battalion models.

Overall, we achieve substantial improvement in terms of model fit – 600, 200 and 131 per cent for SA, OTH and GA contract types, respectively – as measured by the estimation data R^2_{Adj} over the previous effort in (Dertouzos and Garber, 2008). Our improvement comes with some cost of increased p although this is a result of our use of more highly specific geographical market areas[1]. We refrain from direct comparisons to other studies, particularly that of Gibson *et al.* (2011) due to incongruities in the types of responses and units of observation used. Final models for each contract type are provided in Tables VI to VIII.

Validation forecasts

The real test of any quantitative model is how well it performs out-of-sample. For this study, the most recent 15 time periods of data were held out for validation purposes. Each model was forecast out for these 15 periods at a consolidated level (all contract types and battalions combined) and at a detail level (by contract type). Two prediction interval bands are provided, 80 and 95 per cent as analysts vary in how much risk to assume with respect to the certainty of the input independent variables driving the forecast.

Figure 3 is a comprehensive look across all contract types. Within the model data, the prediction values (dark line) track nicely with the actual data (gray line). Tracking is less

Table V.
Summary of fit for the final transformed, lag-1 models with non-significant, non-hereditary terms removed

Response	R^2_{Adj}	R^2_{Pred}	$P > F_0$	p	N
(Reg. Army GA Achieved)$^{1/2}$	0.740	0.730	< 0.001	89	1672
(Reg. Army SA Achieved)$^{1/2}$	0.698	0.679	< 0.001	100	1672
(Reg. Army OTH Achieved)$^{1/2}$	0.807	0.795	< 0.001	98	1672

Battalion	Intercept	β_{x_4}	$\beta_{x_{30}}$	$\beta_{x_{31}}$	$\beta_{x_{32}}$	Lag
BN 1A	6.0013	26.4225	−4.2724	0.0245	−2.0811	0.0028
BN 1B	3.1563	26.4225	0.8758	0.0245	−2.0811	−0.0010
BN 1D	3.5834	26.4225	0.8758	0.0245	−2.0811	−0.0029
BN 1E	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0020
BN 1G	6.5452	26.4225	−7.4793	0.0245	−2.0811	0.0103
BN 1K	4.7563	26.4225	0.8758	0.0245	−2.0811	−0.0008
BN 1N	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0069
BN 1O	4.2881	26.4225	0.8758	0.0245	−2.0811	0.0108
BN 3A	5.6542	26.4225	−3.0809	0.0245	−2.0811	−0.0031
BN 3D	3.9949	26.4225	0.8758	0.0245	−2.0811	0.0070
BN 3E	−1.9996	72.9264	0.8758	0.0245	−2.0811	0.0020
BN 3G	0.8969	26.4225	0.8758	0.0245	−2.0811	0.0106
BN 3H	0.8771	26.4225	0.8758	0.0245	2.9382	0.0046
BN 3J	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0058
BN 3N	0.4258	47.3275	0.8758	0.0245	−2.0811	0.0043
BN 3T	0.3884	26.4225	0.8758	0.0245	−2.0811	0.0199
BN 4C	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0069
BN 4D	1.5753	26.4225	0.8758	0.0358	−2.0811	0.0021
BN 4E	3.3289	26.4225	0.8758	0.0245	−2.0811	0.0060
BN 4G	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0087
BN 4J	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0074
BN 4K	−1.0958	60.5243	0.8758	0.0245	−2.0811	0.0128
BN 4P	2.6972	26.4225	0.8758	0.0245	−2.0811	0.0023
BN 5A	1.3459	26.4225	0.8758	0.0245	−2.0811	0.0094
BN 5C	4.1218	4.2068	0.8758	0.0245	−2.0811	0.0022
BN 5D	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0045
BN 5H	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0060
BN 5I	−0.0926	41.0251	0.8758	0.0245	−2.0811	0.0069
BN 5J	1.4253	26.4225	0.8758	0.0245	−2.0811	0.0151
BN 5K	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0104
BN 5N*	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0058
BN 6F	4.947	26.4225	0.8758	0.0245	−2.0811	−0.0054
BN 6H	2.1978	26.4225	0.8758	0.0245	−2.0811	0.0078
BN 6I	3.7050	26.4225	0.8758	0.0245	−2.0811	−0.0056
BN 6J	1.1932	26.4225	0.8758	0.0437	−2.0811	−0.0005
BN 6K	3.7902	26.4225	0.8758	0.0245	−2.0811	0.0079
BN 6L	2.4804	26.4225	0.8758	0.0245	−2.0811	0.0089
BN 6N	3.491	26.4225	3.7986	0.0245	−2.0811	0.0034

Notes: Coefficient for $\beta_{x_{33}}$ not provided; all less than $|0.001|$; the * indicates baseline

Table VI.
Final GA models for
all Battalions

accurate in the validation data (as one should suspect) but overall not too bad. Figure 4 breaks this data out in the echelon format by contract type. Figure 5 provides a summary of the validation metrics defined for our effort; again, overall these results are very reasonable. We note with particular satisfaction the prediction intervals obtained using linear trend forecasts of the predictors themselves (i.e. "Unknown X"). Only during the very farthest regions of the forecast horizon do we see the actual data e xceed our prediction intervals for both 80 and 90 per cent probability.

We have included the R^2 of each contract type at the aggregate level to compare the estimation models and previous literature for which only Dertouzos and Garber (2008) provide a reliable basis for a comparison; our validation R^2 achieves 530, 170 and 119 per cent relative improvements over the models estimated by Dertouzos and Garber (2008) for

Battalion	Intercept	β_{x_4}	$\beta_{x_{30}}$	$\beta_{x_{31}}$	$\beta_{x_{32}}$	$\beta_{x_{33}}$	Lag
BN 1A	3.7519	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.004
BN 1B	4.8331	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0009
BN 1D	3.3435	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0146
BN 1E	3.0128	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0185
BN 1G	4.8331	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0039
BN 1K	5.7101	15.2018	−0.9503	0.0200	−0.2294	-2.26×10^{-5}	0.0146
BN 1N	3.7381	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0040
BN 1O	4.4737	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0193
BN 3A	4.9531	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0056
BN 3D	5.6615	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0008
BN 3E	0.5804	63.6241	−0.9503	0.0200	−0.2294	3.68×10^{-5}	0.0069
BN 3G	8.3778	15.2018	−0.9503	0.0200	−0.2294	-4.67×10^{-5}	0.0015
BN 3H	4.2080	15.2018	−0.9503	0.0200	3.1698	4.82×10^{-6}	0.0080
BN 3J	4.6361	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0099
BN 3N	0.0017	39.9493	−0.9503	0.0200	−0.2294	4.95×10^{-5}	0.0016
BN 3T	8.0054	15.2018	−4.3841	0.0200	−0.2294	4.82×10^{-6}	0.0029
BN 4C	4.3254	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0066
BN 4D	3.4651	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0146
BN 4E	4.8331	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0006
BN 4G	3.5428	15.2018	−0.9503	0.0298	−0.2294	4.82×10^{-6}	0.0004
BN 4J	4.7788	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0079
BN 4K	0.5943	64.6082	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0086
BN 4P	2.7029	50.8070	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0055
BN 5A	3.4791	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0074
BN 5C	4.2807	15.2018	−0.9503	0.0211	−0.2294	4.82×10^{-6}	0.0078
BN 5D	4.0023	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0060
BN 5H	6.4252	15.2018	−0.9503	0.0114	−0.2294	4.82×10^{-6}	0.0157
BN 5I	4.4255	15.2018	−0.9503	0.0200	−6.0886	4.82×10^{-6}	0.0079
BN 5J	3.2492	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0115
BN 5K	3.3761	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0106
BN 5N*	4.8331	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	−0.0015
BN 6F	4.313	15.2018	−0.9503	0.0200	−0.2294	-1.71×10^{-5}	0.0225
BN 6H	4.2229	15.2018	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0163
BN 6I	4.8073	15.2018	−0.9503	0.0129	−0.2294	4.82×10^{-6}	0.0044
BN 6J	0.9208	40.7777	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0190
BN 6K	0.6777	50.6793	−0.9503	0.0200	−0.2294	4.82×10^{-6}	0.0185
BN 6L	4.0834	15.2018	−0.9503	0.0200	−4.1426	4.82×10^{-6}	0.0160
BN 6N	0.0084	61.7380	−0.9503	0.0200	−0.2294	-7.84×10^{-6}	0.0059

Table VII.
Final OTH models
for all Battalions

Note: The * indicates baseline

SA, OTH and GA contract types, respectively. This is a remarkable feat, especially considering the use of forecast inputs for predictor variables in our models.

Comparative analysis
The causal models developed are based on using multivariate statistical methods to obtain appropriate responses and parsimonious models; these goals were achieved quite well. However, any modeling effort involving many independent variables must answer the question of whether a simpler model would suffice. To this end each model was compared to a naive forecast, an appropriately fit seasonal autoregressive integrated moving average model and a seasonal smoothing method (e.g. the Holt-Winters or Brown's method). Figure 6 plots the

Battalion	QTR 1	QTR 2	QTR 3	QTR 4	β_{x_4}	$\beta_{x_{30}}$	$\beta_{x_{31}}$	$\beta_{x_{33}}$	Lag
BN 1A	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0051
BN 1B	−2.1747	−1.8306	−1.3374	−2.5265	−12.7345	7.3834	0.0107	1.47×10^{-6}	0.0155
BN 1D	2.9588	3.3029	3.7961	2.607	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0335
BN 1E	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0004
BN 1G	1.8122	2.1564	2.2980	1.4604	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0314
BN 1K	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	8.82×10^{-6}	0.0314
BN 1N	3.6116	3.9557	3.5476	3.8873	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0053
BN 1O	−2.2302	−1.2794	−1.3929	−2.582	−12.7345	6.8295	0.0107	1.47×10^{-6}	0.0185
BN 3A	9.5942	9.9383	10.4315	9.2424	−69.8499	0.5186	0.0107	1.47×10^{-6}	−0.0159
BN 3D	−1.6317	−1.2875	−0.7943	−1.9835	−12.7345	0.5186	0.0107	1.30104×10^{-4}	−0.0065
BN 3E	0.6327	0.9769	1.4701	0.2809	−12.7345	5.3432	0.0107	1.47×10^{-6}	0.0126
BN 3G	3.7872	4.1313	4.6245	3.4354	−12.7345	0.5186	0.0023	1.47×10^{-6}	0.0179
BN 3H	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0021
BN 3J	6.9493	7.2935	7.7866	6.5975	−44.7603	0.5186	0.0107	1.47×10^{-6}	0.0015
BN 3N	1.9630	2.3071	2.8003	0.9800	−12.7345	4.6567	0.0107	1.47×10^{-6}	0.0032
BN 3T	3.4437	3.7878	4.281	3.0919	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0221
BN 4C	8.1182	8.4624	8.9555	7.7664	−12.7345	−4.5569	0.0107	1.47×10^{-6}	−0.0141
BN 4D	6.8775	7.2216	7.7148	6.5257	−49.211	0.5186	0.0107	1.47×10^{-6}	−0.0095
BN 4E	3.4288	3.773	4.2662	2.5276	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0082
BN 4G	3.7728	4.1170	4.6102	3.4210	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0131
BN 4J	3.6719	3.4995	4.5093	3.3201	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0076
BN 4K	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0037
BN 4P	6.3981	6.2528	7.2354	6.0463	−25.392	0.5186	0.0107	1.47×10^{-6}	−0.0248
BN 5A	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0126
BN 5C	5.6229	5.9670	6.4602	5.2711	−12.7345	−2.6684	0.0107	1.47×10^{-6}	0.0042
BN 5D	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0052
BN 5H	2.9113	3.2555	3.7487	2.5595	−12.7345	0.5186	0.0192	1.47×10^{-6}	0.0128
BN 5I	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0008
BN 5J	6.2880	6.6322	7.1253	5.9362	−12.7345	−3.0345	0.0107	1.47×10^{-6}	0.0180
BN 5K	3.0773	3.0012	3.9147	2.7255	−12.7345	0.5186	0.0107	1.35×10^{-5}	0.0042
BN 5N*	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.003
BN 6F	3.2325	3.5766	4.0698	2.8806	−12.7345	0.5186	0.0107	1.47×10^{-6}	−0.0046
BN 6H	3.6321	3.9763	4.4695	3.2803	−12.7345	0.5186	0.0107	-1.01×10^{-5}	0.0137
BN 6I	3.5007	3.8449	4.338	3.1489	−12.7345	0.5186	0.0160	1.47×10^{-6}	0.0061
BN 6J	0.2422	0.5864	1.0795	−0.1096	26.5042	0.5186	0.0107	1.47×10^{-6}	0.0145
BN 6K	3.6321	3.9763	4.4695	3.2803	−12.7345	2.642	0.0107	-1.01×10^{-5}	0.0145
BN 6L	3.5736	3.9177	4.4109	3.2218	−12.7345	0.5186	0.0107	1.47×10^{-6}	0.0209
BN 6N	2.6417	2.9858	3.479	2.2899	15.0886	0.5186	0.0107	-2.53×10^{-5}	−0.0131

Note: The * indicates baseline

Table VIII.
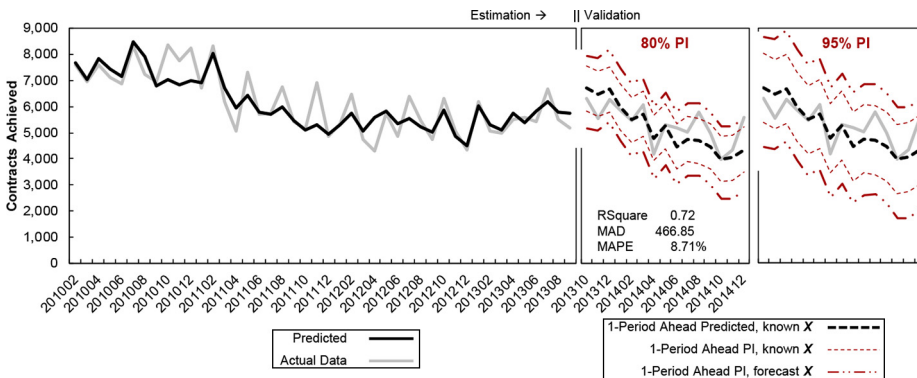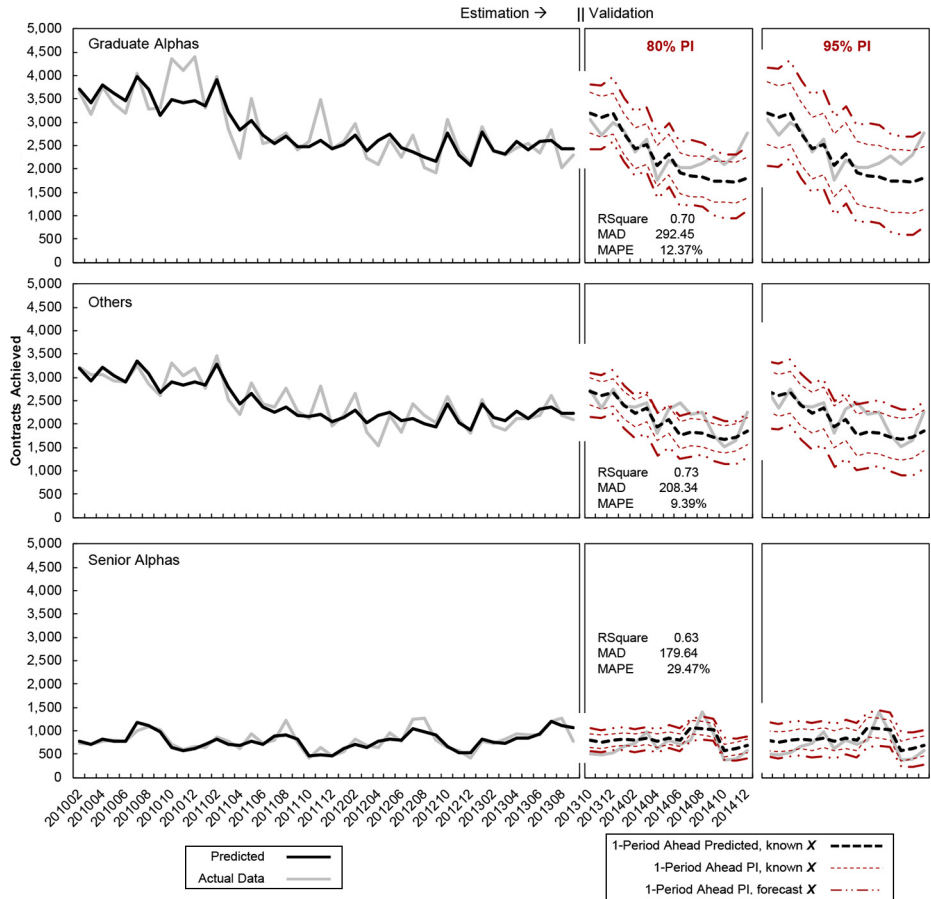Final SA models for
all Battalions

Figure 3.
Contracts achieved and
model forecast for the
total of all contract
types, top-most echelon

Figure 4.
Contracts achieved
and model forecasts
for the three contract
types at top-most
echelon

results for each of the output measures. The legend in each of the sub-figures provide a specification of the univariate model fit; development details are not provided here.

For each of the GA and OTH models, it is quite clear the multivariate models are the preferred approach, realizing of course that these models do already contain a single autoregressive term. For the more seasonal SA response model, the results were not so conclusive as the seasonal time series modeling approaches are comparable options. Overall,
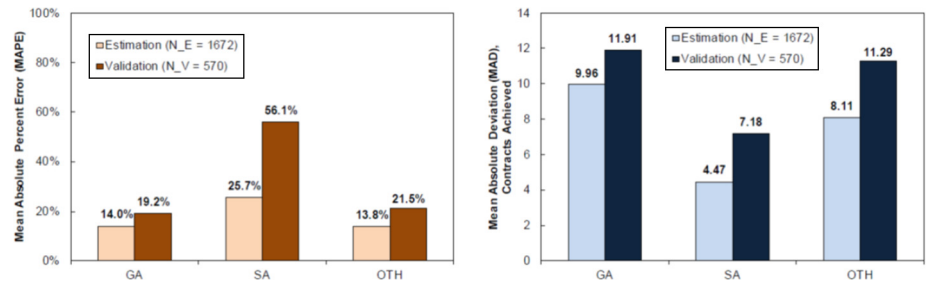
Figure 5.
Summary of model
validation metrics on
the model and hold-
out data sets
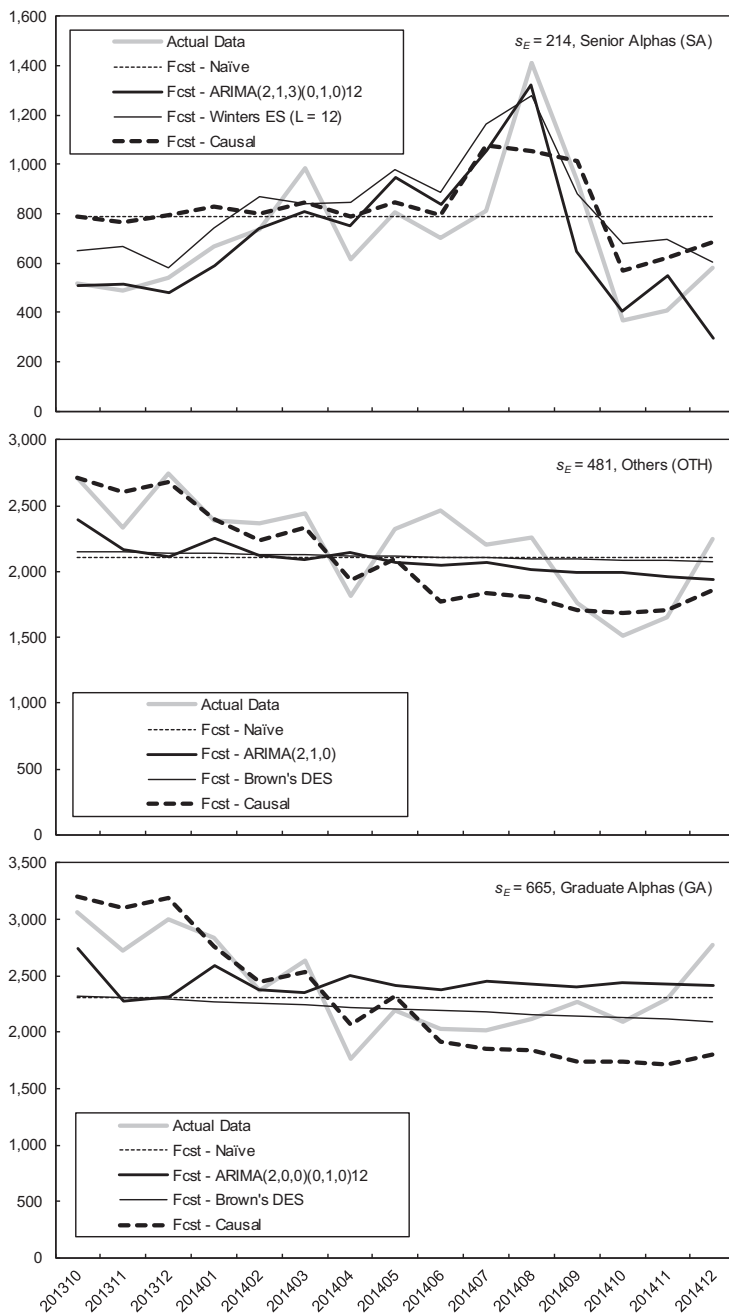
Figure 6.
Forecasts of contracts
achieved, validation
data

the collective result is that our effort to develop multivariate models is indeed rewarded with improved performance.

Conclusion

We have shown, through the use of multiple linear regression aided by increased geographic data specificity through the use of our ZCTA method, mixed stepwise selection methods and PCA, that improvements over previous efforts to model US Army enlistment contract production are possible. Moreover, we have shown that forecasts produced by the multiple linear regression models – which themselves require simple linear forecasts of the predictors – are robust for a relevant forecast horizon of up to 15 months. Indeed, the fit of the forecasts alone constitute remarkable improvements over previous models which did not use validation data and are worth the development effort when compared to simple time series models.

In closing, we must note the unique ability of the multiple regression model in forecasting to consider the effect of inputs which are controlled to a degree by the decision-maker. While the regression model coefficients and standard errors are indeed based only on past data, our models indicated rather high statistical significance of future "controllable" inputs such as recruiting goals and these inputs should be likewise considered in a futuristic sense when the firm is producing forecasts. Only a causal model such as the one afforded by lagged multiple regression affords such an opportunity for exploration. We hope to have successfully informed subsequent discussions on behalf of such a method.

Note

1. Dertouzos and Garber (2008) used 68 variables for each contract type, whereas we used 100, 98 and 89 for SA, OTH and GA models, respectively.

References

Asch, B., Heaton, P. and Savych, B. (2009), "Recruiting minorities: what explains recent trends in the army and navy?", *Technical Report*, RAND Corporation, Santa Monica, CA.

Dertouzos, J. (1985), "Recruiter incentives and enlistment supply", *Technical Report*, RAND Corporation, Santa Monica, CA.

Dertouzos, J. and Garber, S. (2006), "Human resource management and army recruiting: analyses of policy options", *Technical Report*, RAND Corporation, Santa Monica, CA.

Dertouzos, J. and Garber, S. (2008), "Performance evaluation and army recruiting", *Technical Report*, RAND Corporation, Santa Monica, CA.

Flesichmann, M. and Nelson, M. (2014), "Refining recruiting mission allocation using a recruiting market index", *Presentation to the Army Operations Research Symposium*.

Gibson, J., Hermida, R., Luchman, J., Griepentrog, B. and Marsh, S. (2011), "ZIP Code Valuation Study Technical Report", *Technical Report*, Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia.

Gibson, J., Luchman, J., Griepentrog, B., Marsh, S., Zucker, A. and Boehmer, M. (2009), "ZIP code valuation study technical report: predicting army accessions", *Technical Report*, Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia.

Howard, R. and Abbas, A. (2015), *Foundations of Decision Analysis*, Prentice Hall, Upper Saddle River, New Jersey.

McDonald, J. (2016), "Analysis and modeling of u.s. army recruiting markets", Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, OH.

Montgomery, D., Jennings, C. and Kulachi, M. (2015), *Introduction to Time Series Analysis and Forecasting*, 2nd ed., John Wiley and Sons, Hoboken, New Jersey.

Montgomery, D., Peck, E. and Vining, G. (2012), *Introduction to Linear Regression Analysis*, 5th ed., John Wiley and Sons, Hoboken, New Jersey.

Murray, M. and McDonald, L. (1999), "Recent recruiting trends and their implications for models of enlistment supply", *Technical Report*, RAND Corporation, Santa Monica, California.

U.S. Census Bureau (2015), "*2010 ZCTA to County Relationship File*", available at: www2.census.gov/geo/docs/maps-data/data/rel/zcta_county_rel_10.txt, (accessed 24 September 2015).

Wackerly, D., Mendenhall, W. and Schaeffer, R. (2008), *Mathematical Statistics with Applications*, 7th ed., Brooks-Cole Cengage, Belmont, CA.

Waddell, S. (2005), *History of the Military Art since 1914*, Pearson Custom Publishing, West Point, New York.

Warner, J., Simon, C. and Payne, D. (2001), "Enlistment supply in the 1990's: a study of the navy college fund and other enlistment incentive programs", *Technical Report*, Defense Manpower Data Center, JAMRS Division, Arlington, Virginia.

Corresponding author
Raymond R. Hill can be contacted at: rayrhill@gmail.com