Journal of Forest Economics 18 (2012) 76–89

Contents lists available at SciVerse ScienceDirect

Journal of Forest Economics

journal homepage: www.elsevier.de/jfe

An econometric model for *ex ante* prediction of wildfire suppression costs

Jonathan Yoder[a], Krista Gebert[b,*]

[a] *School of Economic Sciences, Washington State University, United States*
[b] *Northern Region, USDA Forest Service, United States*

ARTICLE INFO

Article history:
Received 6 November 2010
Accepted 19 October 2011

JEL classification:
Q20
C13

Keywords:
Wildfire suppression costs
Bivariate truncated regression

ABSTRACT

This paper develops an econometric model that can provide predictions of fire suppression costs (per acre and in total) for a given large fire before final fire acreage is known. The model jointly estimates cost per acre and acreage equations via Maximum Likelihood, accounting for sample truncation based on final fire size. Formulas and results are shown for predictions of costs and fire size for wildfires in general, and for large fires in particular. Marginal effects of explanatory variables on cost and acreage are discussed. The distribution of these model predictions illustrates the importance of accounting for sample truncation when generating predicted outcomes based on *ex ante* information.

Introduction

Between 1980 and 2002, wildfires larger than 300 acres amounted to about 1.4 percent of all reported wildfires, but accounted for about 94 percent of all suppression expenditures. Large fires are responsible for the bulk of fire suppression expenditures (Strategic Issues Panel on Fire Suppression Costs, 2004). Understanding the factors that influence suppression costs and size of large fires is therefore important for both strategic fire planning and on-site fire management decisions. This paper contributes to the existing literature and modeling approaches by developing and estimating a bivariate econometric model of cost per acre and fire size that can provide statistically consistent and relatively precise predictions of fire suppression costs (per acre and in total) for a given large fire before final fire acreage is known.

* Corresponding author. Tel.: +1 509 335 8596; fax: +1 509 335 1173.
E-mail address: yoder@wsu.edu (J. Yoder).

1104-6899/$ – see front matter © 2011 Department of Forest Economics, SLU Umeå, Sweden. Published by Elsevier GmbH. All rights reserved.
doi:10.1016/j.jfe.2011.10.003

Several papers have been published that focus on estimating costs and/or acreage of individual wildland fires. The most recent and relevant include Butry et al. (2008), who estimate log-linear regression models for fire acreage based on fire characteristics, climate/weather, management and mitigation efforts, and other factors. They break their sample into two subsamples (fires > 1000 acres and fires < 1000 acres) and estimate the same regression equation on each sample separately to allow flexibility to account for structural differences between large and small fire regimes. However, they do not statistically account for truncation of their subsamples (limited to 1000 acres or larger), which may introduce bias and inconsistency in parameter estimates and model predictions.

Gebert et al. (2007) estimate suppression costs per acre for individual wildland fires of 100 acres or more. The paper focuses specifically on costs per acre, and does not estimate a fire size equation to address sample truncation based on fire size. Currently, models such as this are being used for forecasting during the fire, but with reliance on ad hoc estimates of final fire acreage. Holmes et al. (2008) estimate a set of models for fire size based on an extreme value threshold model based on a generalized Pareto distribution to allow for relatively heavy tailed distributions. One distinction of this approach is that it calls for the explicit selection of a fire size threshold for including an observation in the sample for estimation.

The primary contribution of this present paper is the development and estimation of a model of suppression costs for individual fires that accounts for sample truncation inherent in our data, and is useable for forecasting prior to knowing final fire characteristics.

We estimate acreage and cost per acre equations as a bivariate system of equations. This allows us to better utilize information about the relationship between acreage and costs, thereby improving forecasting precision while allowing early cost predictions (prior to knowing acreage). Addressing sample truncation based on acreage addresses potential statistical inconsistency and bias in parameter estimates and predictions, for both acreage and cost equations, that would likely exist if truncation were ignored. We use new data from the Department of Interior for 2004 through 2009, for a total of 2061 available observations on fires of 300 acres or more.

Although the modeling framework of Holmes et al. (2008) is a reasonable approach, we rely on lognormal disturbances for modeling because it provide a more stable and manageable framework for joint estimation of costs and acreage with a relatively large number of covariates. We contend that this distributional assumption is a reasonable approximation and worth the benefits in terms of practical estimation and forecasting. We also found in preliminary analysis that applying the threshold selection approach used by Holmes et al. (2008) would call for substantial data and information loss (up to 95% of our observations).

The next section provides the theoretical foundation for the empirical model, followed by data descriptions, estimation results, cost and fire size prediction summaries, and a conclusion.

Model and estimation

As motivation for the bivariate regression equations for cost per acre and fire acreage, suppose a fire suppression manager allocates resources to balance suppression costs against wildfire damage losses. Consider two types of suppression inputs: s_a is productive for limiting acreage, and s_d is productive for limiting damage per acre, with constant marginal costs r_a and r_d, respectively. Total suppression costs are $T = r_a s_a + r_d s_d$. Fire size $A(s_a, \boldsymbol{x}_a, \varepsilon_a)$, is a decreasing function of s_a, and is also affected by exogenous factors $\boldsymbol{x}_a$ and a random disturbance ε_a. Per acre suppression costs are therefore $C(s_a, s_d) = T/A(s_a, \boldsymbol{x}_a, \varepsilon_a)$. Damage per acre is $D(s_d, \boldsymbol{x}_d, \varepsilon_d)$, which is a decreasing function of s_d, and is also a function of exogenous factors $\mathbf{x}_d$ and a random disturbance ε_d. Total losses from the fire are then $L(\cdot) = D(s_d, \boldsymbol{x}_d, \varepsilon_d)A(s_a, \boldsymbol{x}_a, \varepsilon_a)$. The random disturbances are unobservable in the data, assumed uncorrelated with exogenous variables, and treated as i.i.d. random disturbances for *ex post* observation and estimation. This theoretical model formulation is designed to illustrate how endogenous suppression

allocation will induce a close relationship between per acre costs and acreage that supports joint estimation.[1]

Assume that the fire manager allocates suppression resources to maximize utility over suppression costs and wildfire damage. The unconstrained maximization problem can be characterized as

$$\max_{s_a,s_d} U\{T(s_a, s_d),\ L(s_a, s_d)\}, \tag{1}$$

where exogenous variables and disturbances are omitted for clarity but will be reintroduced shortly. The necessary conditions for a maximum are

$$\frac{\partial U}{\partial s_d} = U_T \frac{\partial T}{\partial s_d} + U_L A \frac{\partial D}{\partial s_d} = 0 \tag{2}$$

and

$$\frac{\partial U}{\partial s_a} = U_T \frac{\partial T}{\partial s_a} + U_L D \frac{\partial A}{\partial s_a} = 0, \tag{3}$$

where U_T and U_L are the marginal utility of suppression costs and losses, respectively. They are both assumed to be negative, but the working environment of a fire manager may induce differences in the marginal valuation of a dollar of suppression and a dollar loss in values at risk. Sufficient second-order curvature conditions are assumed to hold for a maximum.

Consider Eq. (3) further. Recall that $T = AC$, so it follows that $\partial T/\partial s_a = A(\partial C/\partial s_a) + (\partial A/\partial s_a)C$. Eq. (3) can therefore be written as

$$U_T \left(A \frac{\partial C}{\partial s_a} + \frac{\partial A}{\partial s_a} C \right) + U_L D \frac{\partial A}{\partial s_a} = 0. \tag{4}$$

Rearranging and simplifying Eq. (4) provides

$$\frac{\partial C}{\partial A} = -\left(\frac{C}{A} + \frac{U_L}{U_T} \frac{D}{A} \right) < 0. \tag{5}$$

Thus, we would expect to find a negative relationship between costs per acre and fire size. This is indeed the case in most wildfire data: based on raw correlations, cost per acre tends to be substantially lower for larger fires.[2]

Assuming necessary and sufficient conditions hold for a maximum and the implicit function theorem holds, it can be shown that optimal demand for both types of suppression is a function of all exogenous observed and unobserved factors: $\mathbf{s}^* = [s_a(\mathbf{x}, \boldsymbol{\varepsilon}), s_d(\mathbf{x}, \boldsymbol{\varepsilon})]'$, where $\mathbf{x} = \mathbf{r} \cup \boldsymbol{x}_d \cup \boldsymbol{x}_a$ is the union of all exogenous variables and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_a & \varepsilon_d \end{bmatrix}$. Substituting the suppression demand function back into the C and A provides the indirect cost per acre and acreage functions as two outcomes of a fire:

$$\begin{aligned} C^*(\mathbf{x}, \boldsymbol{\varepsilon}) &= \frac{\mathbf{r}'\mathbf{s}(\mathbf{x}, \boldsymbol{\varepsilon})}{A^*(s(\mathbf{x}, \boldsymbol{\varepsilon}), \boldsymbol{x}_a)}, \\ A^*(\mathbf{x}, \boldsymbol{\varepsilon}) &= A^*(s(\mathbf{x}, \boldsymbol{\varepsilon}), \boldsymbol{x}_a). \end{aligned} \tag{6}$$

We do not have data on suppression effort s^*, and so cannot directly estimate suppression demand functions. However, given that suppression demand is a function of exogenous factors, indirect

[1] Two inputs are necessary to mathematically identify the choice between reducing damage per acre and reducing acreage. The separability in production between the two inputs is somewhat restrictive, but provides a very clear foundation for illustrating how endogenous suppression allocation induces a tight relationship between costs per acre and fire size.

[2] This often observed negative correlation between fire size and cost per acre have motivated several others in the literature (including the authors of this paper – see Gebert et al., 2007) to hypothesize that fire size should enter directly into the cost/acre equation and/or that cost/acre should enter into the acreage equation. However, as our theoretical model demonstrates, optimal fire suppression investment may give rise to this correlation even when no simultaneous causality exists between the two outcomes. Further, we think that there is no reason to suspect otherwise. Suppression costs follow from suppression effort; why should fire size (a physical outcome) affect suppression costs (an economic outcome) except through its relation to chosen suppression effort? Why should costs per acre (economic) affect fire size (physical) except through its relation to chosen suppression effort? We cannot conceive of a reason for either case, and we therefore maintain the view that the two outcomes are correlated but do not directly (causally) affect each other.

cost/acre and acreage equations can be estimated as a function of exogenous factors through implicit suppression demand (this is a standard duality theory result). Finally, note that endogenous cost per acre and acreage are each a function of both disturbances, and so, again, we would expect to find a relationship in the (implicitly compound) disturbance process in estimated versions of these equations.

The statistical distribution of disturbances in empirical work on fire size and costs approximate a lognormal. A linear approximation of this two-equation system $[C^*(\mathbf{x}, \boldsymbol{\varepsilon}), A^*(\mathbf{x}, \boldsymbol{\varepsilon})]$ for our estimation purposes is therefore characterized as

$$\begin{aligned} \mathbf{c} &= \mathbf{X}\boldsymbol{\beta}_c + v_c \\ \mathbf{a} &= \mathbf{X}\boldsymbol{\beta}_a + v_a, \end{aligned} \tag{7}$$

where $\mathbf{c} = \ln(\mathbf{C})$ and $\mathbf{a} = \ln(\mathbf{A})$ are log-transformed vectors of cost per acre and acreage, and $v_c = f_c(\boldsymbol{\varepsilon})$ and $v_a = f_a(\boldsymbol{\varepsilon})$ are assumed normally distributed, related to $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_a & \varepsilon_d \end{bmatrix}$, and likely correlated through $s^* = s(\mathbf{x}, \boldsymbol{\varepsilon})$ for a given fire.[3]

The Maximum Likelihood counterpart for the theoretical model is now developed. Because the dataset includes data on large fires only (e.g. >300 acres), the disturbances v_a are also truncated (v_a cannot be less than $\ln(300) - \mathbf{X}\boldsymbol{\beta}_a$), and this may lead to biased and inconsistent parameter estimates if not accounted for.[4] Given the log transformation of A and C and a minimum fire size, the conditional distribution of the reduced form regression disturbances $\mathbf{V} = [v_a\, v_c]$ can be represented for a given observation as (observation index i suppressed)[5]:

$$f(v_c, v_a | a \geq k) = \frac{f(v_c, v_a)}{\Pr(a \geq k)}, \tag{8}$$

where $k = \ln(K)$ is the logarithm of the minimum fire size K. The condition $a \geq k$ implies $a - \mathbf{x}\boldsymbol{\beta}_a \geq k - \mathbf{x}\boldsymbol{\beta}_a$, where $\mathbf{x} = \mathbf{X}_i$ represents observation $i \in N$ on each variable in the $(N \times J)$ matrix $\mathbf{X}$, where N is the total number of observations and J is the number of variables (constant included). The joint distribution $f(v_c, v_a)$ is

$$\begin{aligned} f(v_c, v_a) &= \frac{1}{AC} \frac{1}{2\pi\sigma_a\sigma_c\sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho^2)}\right), \\ \text{where, } z &= \left(\frac{a - \mathbf{x}\boldsymbol{\beta}_a}{\sigma_a}\right)^2 + \left(\frac{c - \mathbf{x}\boldsymbol{\beta}_c}{\sigma_c}\right)^2 - \left(\frac{2\rho(a - \mathbf{x}\boldsymbol{\beta}_a)(c - \mathbf{x}\boldsymbol{\beta}_c)}{\sigma_a\sigma_c}\right), \\ \text{and } \Pr(a \geq k) &= \Pr(v_a \geq k - \mathbf{x}\boldsymbol{\beta}_a) = 1 - \Phi\left(\frac{k - \mathbf{x}\boldsymbol{\beta}_a}{\sigma_a}\right). \end{aligned} \tag{9}$$

The associated log-likelihood function is

$$\ln L = -\sum\nolimits_n \left[(a+c) + \ln\left(2\pi\sigma_a\sigma_c\sqrt{1-\rho^2}\right) + \left(\frac{z}{2(1-\rho^2)}\right) + \ln\left(1 - \Phi\left(\frac{k - \mathbf{x}\boldsymbol{\beta}_a}{\sigma_a}\right)\right)\right]. \tag{10}$$

Parameter estimates for $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_c, \sigma_a, \sigma_c, \rho)$ are the values that jointly maximize the log-likelihood function (10).

[3] Cobb–Douglas functional forms for production along with specific utility function forms can exactly imply this linear regression structure for costs per acre and acreage.

[4] See Greene (2008) for a more in-depth description of truncation and Mostafa and Mahmoud (1964) for the bivariate log-normal distribution.

[5] See Lien and Balakrishnan (2006) for theory underlying this type of model.

Model prediction calculations

Following Lien (1985), the expectations of the untransformed original acreage and cost variables are the following for observation $i \in N$:

$$E[A_i] = \exp\left(\boldsymbol{x}_i\boldsymbol{\beta}_a + \frac{\sigma_a^2}{2}\right), \tag{11}$$

$$E[C_i] = \exp\left(\boldsymbol{x}_i\boldsymbol{\beta}_c + \frac{\sigma_c^2}{2}\right), \tag{12}$$

The medians of $Z_i \in (A_i, C_i)$ are $M[Z_i] = \exp(\boldsymbol{x}_i\boldsymbol{\beta}_z)$ (Mostafa and Mahmoud, 1964). The variance for expected acreage is calculated using the Delta method:

$$\widehat{V[E[A_i]]} = \mathbf{G}\widehat{\mathbf{V}[\hat{\boldsymbol{\theta}}]}\mathbf{G}' \tag{13}$$

where $\boldsymbol{\theta}_{1\times(m+1)} = (\boldsymbol{\beta}_z, \sigma_z)$ is the vector of m parameters in $\boldsymbol{\beta}_z$ as well as σ_z, $Z \in (A, C)$ and $z \in (a, c)$, $\widehat{\mathbf{V}[\hat{\boldsymbol{\theta}}]}_{(m+1)\times(m+1)}$ is the estimated covariance of the estimates $\hat{\boldsymbol{\theta}}$, and

$$\mathbf{G}_{1\times(m+1)} = \left.\frac{\partial E[Z(\boldsymbol{\theta}, \mathbf{X})]}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{14}$$

For our specific case with $E[Z_i]$ defined by Eqs. (11) and (12), the gradient $\mathbf{G}$ is

$$\mathbf{G} = \begin{bmatrix} E[Z_i]x_{i1} & E[Z_i]x_{i2} & \cdots & E[Z_i]x_{ij} & E[Z_i] \end{bmatrix}. \tag{15}$$

The expected value of total expenditures is[6]:

$$E[A_i \cdot C_i] = \exp\left(\boldsymbol{x}_i\boldsymbol{\beta}_c + \boldsymbol{x}_i\boldsymbol{\beta}_a + \frac{(\sigma_c^2 + \sigma_a^2)}{2} + \rho\sigma_a\sigma_c\right) = E[A_i]\cdot E[C_i]\cdot\exp(\rho\sigma_a\sigma_c). \tag{16}$$

Given that the truncation on A has been accounted for in parameter estimation, replacing ($\boldsymbol{\beta}_a$, $\boldsymbol{\beta}_c$, σ_a, σ_c, ρ) in Eqs. (11)–(16) provides consistent predictions for acreage, costs/acre, and their variances, and should be the basis for *ex ante* prediction and inference for acreage and cost/acre.

If cost and acreage predictions are desired and estimated only for fires that reach size K or greater, the conditional expectations below should be used for prediction:

$$E[A_i|A_i > K] = E[A_i]\cdot\frac{\Phi(\sigma_a - (k - \boldsymbol{x}_i\boldsymbol{\beta}_a)/\sigma_a)}{\Phi(-(k - \boldsymbol{x}_i\boldsymbol{\beta}_a)/\sigma_a)}, \tag{17}$$

$$E[C_i|A_i > K] = E[C_i]\cdot\frac{\Phi(\rho\sigma_c - (k - \boldsymbol{x}_i\boldsymbol{\beta}_a)/\sigma_a)}{\Phi(-(k - \boldsymbol{x}_i\boldsymbol{\beta}_a)/\sigma_a)}, \tag{18}$$

$$E[A_i \cdot C_i|A_i > K] = E[A_i]\cdot E[C_i]\cdot\rho\sigma_c\sigma_a\cdot\frac{\Phi((\sigma^2 + \rho\sigma_c\sigma_a - (k - \boldsymbol{x}_i\boldsymbol{\beta}_a))/\sigma_a)}{\Phi(-(k - \boldsymbol{x}_i\boldsymbol{\beta}_a)/\sigma_a)}, \tag{19}$$

and $\Phi(\cdot)$ is the standard normal cumulative density.[7]

[6] Variances can be calculated via the Delta method in this case as well.

[7] If acreage A and cost per acre C are both conditionally lognormal, then expenditures $E = AC$ is also lognormal. Eqs. (7)–(12) and the estimation process would still be valid for estimating expenditures (CA) with truncation based on A simply by replacing the variable C with the total expenditures variable CA and implementing maximum likelihood based on $\ln(AC) = (a + c)$ as the dependent variable. This might be a more direct approach for predicting CA, it complicates estimation of C alone and does not save on degrees of freedom. On a related note, Lien and Balakrishnan (2006) provide a general treatment of the conditional distribution of multiplicatively constrained lognormal distributions such as would be the case if sample truncation were based on total expenditures, as in a sample that included fires such that $A \cdot C \geq K$.

Marginal effects and percentage effects

The dependent variables in the regressions are the logarithm of C and A, and all regressors are transformed in some way prior to estimation. Therefore, the parameters on continuous variables take the general form $\beta_j = \partial \ln(y_i)/\partial g(x_{ij})$, where $y_i \in (C_i, A_i)$, $g(\cdot)$ is regressor transformation function, $i \in N$ is the observation index and $j \in J$ is a regressor index. The marginal relationships between an exogenous variable and the dependent variable are then $\partial y_i/\partial x_i = y_i(\partial \ln(y_i)/\partial g(x_{ij}))(\partial g(x_{ij})/\partial x_{ij}) = y_i\beta_j(\partial g(x_{ij})/\partial x_{ij})$. The parameter estimates on all continuous variables except *Energy Release* and the sine and cosine variables are log transformations. For these log-transformed variables, the parameter estimates represent elasticities: $\beta_j = \partial \ln(y_i)/\partial \ln(x_{ij}) = (\partial y_i/\partial x_{ij})(x_{ij}/y_i)$ for a given regressor x_j, and the marginal effects are $\partial y_i/\partial x_{ij} = \beta_j(y_i/x_{ij})$. The most complicated regressor transformations apply to variables with cyclical effects, such as month and aspect, which are transformed into cosine and sine waves. The *month* transformation involves translating each of the 12 months into radians and then applying *sine* and *cosine* functions. So, if $m = month$, then $h(x_{ij}) = \beta_1 \sin(x_{ij}) + \beta_2 \cos(x_{ij})$, where $x_{ij} = 2\pi((m_i - 1)/12)$ and *month* m_i is a categorical variable ranging from 1 (January) to 12 (December). The marginal effect of a change in month is $\partial y_i/\partial m_i = (2\pi y_i/12)(\beta_1 \cos(2\pi(m_i - 1)/12) + \beta_2 \sin(2\pi(m_i - 1)/12))$. This marginal effect may switch from positive to negative over the course of a year depending on m_i. The variable *Aspect* is similarly transformed into radians and then into *sine* and *cosine* functions.

The marginal effects and elasticities of a continuous variable on $E[A_i]$, $E[C_i]$, and $E[A_i \cdot C_i]$ are relatively straightforward to calculate. For example, based on Eq. (16), the marginal effect of x_j on $E[A_i \cdot C_i]$ is $\partial E[A_i \cdot C_i]/\partial x_j = (\beta_{aj} + \beta_{cj}) \cdot E[A_i \cdot C_i]$, which would be calculated by replacing all of the parameters on the right hand side of Eq. (16) with their estimated counterparts. The associated percentage effect of this variable x_j on $E[A_i \cdot C_i]$ would be $(\beta_{aj} + \beta_{cj})$, and if $x_j = \ln(X_j)$, then $(\beta_{aj} + \beta_{cj})$ would represent an elasticity of $E[A_i \cdot C_i]$ with respect to X_j. The marginal effects of a change in a variable on the *conditional* expectations of A, C, and $A \cdot C$ (as characterized in Eqs. (17)–(19)) are more complicated functions involving the standard normal CDF and PDF.

For indicator (dummy) variables, the regression parameter estimates do not represent a percentage change in y_i due to the category represented by the indicator variable. Rather, an unbiased and consistent estimator of the percentage change from the base case of Y_i due to an indicator variable d is[8]:

$$\hat{p}_d = 100\left(\exp\left(\hat{\beta}_d - \frac{1}{2}\hat{v}[\hat{\beta}_d]\right) - 1\right). \tag{20}$$

The total estimated difference in the dependent variable y evaluated at some vector $\mathbf{x}$ and the base case (with $d = 0$) would then be $\hat{y}(\mathbf{x}, d = 0) \cdot \hat{p}_d$. Similarly, the percentage difference in y_i between two non-base categories represented by dummy variables d_1 and d_2 can be can be calculated as

$$\hat{p}_{d12} = 100\left(\exp\left(\hat{\theta}_{21} - \frac{1}{2}\hat{v}[\hat{\theta}_{21}]\right) - 1\right), \tag{21}$$

where $\hat{\theta}_{21} = (\hat{\beta}_{d2} - \hat{\beta}_{d1})$ and $\widehat{v}[\hat{\theta}_{21}] = \widehat{Var}[\hat{\beta}_{d_2}] + \widehat{Var}[\hat{\beta}_{d_1}] + 2\widehat{Cov}[\hat{\beta}_{d_1}, \hat{\beta}_{d_2}]$.[9]

Percentage effects of a difference in a categorical variable from the base case on total expenditures $E[A_i \cdot C_i]$ can be calculated using Eq. (21) using $\hat{\theta}_{ac} = (\hat{\beta}_{ad} + \hat{\beta}_{cd})$, where $\hat{\beta}_{ad}$ and $\hat{\beta}_{cd}$ are the coefficients on parameters associated with a dummy variable d in the acreage and cost equation respectively.[10]

[8] See van Garderen and Shah (2002), who also provide an estimator for the variance of $\hat{p}$.

[9] Consider two observations on Y in a model with just two dummy variables: observation y^1 has dummy variable 1 equal to 1 and dummy variable 2 equal to zero, and y^2 has dummy variable 2 equal to 1 and dummy variable 1 equal to zero, so that $y^1 = \alpha + \beta_1 d_1 + \varepsilon$ and $y^2 = \alpha + \beta_2 d_2 + \varepsilon$. The percent difference between y^2 and y^1 is $(y^2 - y^1)/y^1 = (\exp(\alpha + \beta_2 + \varepsilon_1) - \exp(\alpha + \beta_1 + \varepsilon_2))/\exp(\alpha + \beta_1 + \varepsilon_1) = \exp(\beta_2 - \beta_1)\exp(\varepsilon_2 - \varepsilon_1) - 1$. Taking expectations, applying Kennedy's (1981) bias correction, and substituting $\hat{\theta}_{21} = (\hat{\beta}_{d2} - \hat{\beta}_{d1})$ provide $E[(y^2 - y^1)/y^1] = \exp((\hat{\theta}_{21}) + (1/2)\hat{v}[\hat{\theta}_{21}]) - 1$. Multiply this by 100 to get a percentage value.

[10] Comparing the effects of two non-base categories on total expenditures using parameter estimates both equations is a straightforward but tedious extension of this calculation.

Data

The fires used in this analysis come from a database maintained by the Rocky Mountain Research Station (RMRS). This database contains expenditure data and fire characteristic information for individual large (300 + acre) wildfires reported by the USDA Forest Service and the Department of Interior (DOI). The data on large DOI fires includes the years 2004 onward. Expenditures include federal expenditures obtained from the financial systems of the USDA Forest Service and the DOI and do not include state expenditures, except those already accounted for in the federal expenditure databases. Therefore, the expenditures on any given wildfire are the sum of the federal expenditures on that particular fire.

Fire characteristic data in the RMRS database come either directly from the information reported in the fire occurrence databases of the federal land management agencies or are calculated using the reported latitude and longitude of the ignition point of the fire. The federal fire occurrence databases include: (1) the National Interagency Fire Management Integrated Database (NIFMID) – Forest Service, (2) the Wildland Fire Management Information system (WFMI) – Bureau of Land Management, Bureau of Indian Affairs, and National Park Service, and (3) the Fire Management Information System (FMIS) – Fish and Wildlife Service. See Gebert et al., 2007 for a fuller description of the data and data collection process.

We do not have data for marginal suppression costs *r*. We utilize annual and USFS district dummy variable in our regressions in an attempt to capture aggregate but unobservable differences across years and regions, including variation in labor and rental rates for suppression equipment.

We restrict our analysis to fiscal years 2004 through fiscal year 2009 (the most current information during analysis). DOI expenditure data for individual fires is difficult and time consuming to collect given the different accounting systems and rules used by each of the DOI agencies. In fiscal year 2004, the FireCode system was implemented "to standardize fire incident financial coding for fire suppression and subsequent emergency stabilization . . . to provide the capability to effectively track and compile the full cost of a multi-jurisdictional fire suppression effort" (see USDI, 2011). Therefore, from fiscal year 2004 on, the expenditure data collected by RMRS included all federal expenditures, whereas previously it had only included Forest Service expenditures. Additionally, we also restricted our analysis to fires 300 acres or more in size, which is the size generally associated with fires that have escaped initial attack. Moreover, equations developed to estimate expenditures are more likely to be used once a fire escapes initial attack, when it becomes subject to more oversight and cost becomes more of an issue due to larger acreages and longer durations.

Our final dataset consists of 2061 DOI fires. The dependent and independent variables used in this analysis are defined in Table 1, with summary statistics for our dataset provided in Table 2.

Results

This section reports regression results, including some detailed discussion of selected outcomes for interpretive illustration. We also provide a brief comparison of the bivariate regression performance with results from a univariate regression approach. We then provide cost per acre, acreage, and expenditure forecasts based on the bivariate truncated regression.

Regression estimates

Table 3 provides regression parameter estimates. A limited set of the estimated regression relationships are interpreted here for illustration. First consider the effect of *Elevation*. The estimated change in fire acreage in response to a one percent higher elevation is −0.0456% (though not statistically significantly so, with $p = 0.885$), and costs per acre increase by an estimated 0.2714% ($p < 0.075$).[11] The variable *Energy Release* ranges from 1 to 100, and so the parameter represents a percent change in the dependent variable with respect to a one unit change in *Energy Release*. In this case, acreage increases

[11] *p*-Values are omitted in the tables to conserve space.

Table 1
Independent variables used in development of regression equations.

Fire characteristics	Variable definition	Source
Fire environment		
sin(*Aspect*), cos(*Aspect*)	Sine and cosine of aspect at point of origin in 45 degree increments	Fire Occurrence Databases
Slope	ln(Slope percent at point of origin)	Fire Occurrence Databases
Elevation	ln(Elevation at point of origin)	Fire Occurrence Databases
Latitude	Decimal degrees	Fire Occurrence Database
fuel i	Dummy variables representing fuel type at point of origin. Grass = NFDRS fuel model A, L, S, C, T, N; Brush = NFDRS fuel model F, Q; Slash = NFDRS fuel model J, K, I; Timber = NFDRS fuel model H, R, E, P, U, G; brush 4 (reference category) = NFDRS fuel model B, O.	Fire Occurrence Databases
Energy release	Energy release component (ERC) calculated from ignition point using nearest weather station information (cumulative frequency)	Calculated
values at risk		
ln(*distance*)	Natural log of distance from ignition to nearest census designated place	Calculated
ln(*housing* 20)	Natural log of total housing value in 20 mile radius from point of origin (census data)/100,000	Calculated
Wilderness	Dummy variable = 1 if originated in a wilderness area, zero otherwise	Calculated
Other		
sin(*Month*), cos(*Month*)	Sine and cosine of month/year translated into radians	Calculated
Natural cause	Dummy variable = 1 if natural, 0 if human caused	Fire occurrence database
year	Dummy variables representing fiscal year	Fire occurrence database
region i	Dummy variables for Geographic Area Coordination Center (DOI), which represent distinct management regions.	Fire occurrence database

by 0.0326% (p-value <0.001), and cost/acre decreases by 0.0041% with a one unit change in *Energy Release* (though not with statistical significance: $p = 0.17$).

Wilderness is a dummy variable representing fires that started in designated Wilderness areas. Based on Eq. (20), fires in wilderness areas are an estimated 904% larger ($p < 0.001$), and costs/acre are about 46% lower than fires outside of Wilderness designations ($p < 0.001$). *Year*, *region*, and *fuel* type are represented by dummy variables as well, with 2003, *region* 1, and *fuel type* 4 (grass) as the base case. Thus, the coefficients on the *region* 2, for example, represent the difference between *region* 1 and *region* 2. Again using Eq. (20), region 2 tends to have fires that are 86% smaller than *region* 1, and cost/acre 5.6% smaller than *region* 1, though this last estimate is not statistically different from zero at conventional confidence levels. Calculating the percent difference between two non-base cases relies on the content of footnote 9. For example, acreage of fires categorized as *fuel type* 5 (timber/slash) are an estimated 69% smaller ($p < 0.001$) than acreage of fires categorized as *fuel type* 2 (brush_4 [chaparral]), but cost/acre are estimated to be 126% higher ($p = 0.040$) for timber/slash fires than chaparral fires (the covariances between the parameters necessary to calculate these two values are 0.00107 and 0.00093, respectively).

The effects of explanatory variables on total expenditures can also be calculated. For example, the percentage change in predicted expenditures $E[A_i \cdot C_i]$ with respect to a change in *Energy Release* component is $(0.0326 - 0.0041) = 0.0285$ ($p < 0.001$). The elasticity of with respect to a change in elevation is $(-0.0456 + 0.2714) = 0.2258$, though this is not statistically different from zero ($p = 0.324$). Finally, a fire started in a wilderness area is associated with 440% higher expenditures than otherwise ($p < 0.001$), based on Eq. (21).

Importantly, note that the estimate for the correlation between errors is $\rho = -0.51$ ($p < 0.0001$). This negative correlation suggests that, after all included explanatory variables have been accounted for, the unobserved component of costs per acre are high when acreage is low, and vice versa. This is consistent with our theory (Eq. (5)) for the effects of unobserved (random) factors, and widely observed negative correlation between cost per acre and fire size more generally. It also indicates that selecting

Table 2
Summary statistics.

Variable	$N = 2061$			
	Mean	S.D.	Min	Max
A	10,262	45,449	300	1,000,000
C	204.66	673	1.04	19,535
ln(*elevation*)	1.27	0.67	0.00	2.30
Latitude	40.74	8.59	2.84	68.26
ln(*housing* 20)	14.99	10.49	−9.21	25.93
Energy Release	80.31	18.43	1.00	100.00
ln(*distance*)	2.45	0.94	−2.14	6.38
ln(*slope*)	0.26	0.43	0.00	1.61
cos(*Aspect*)	0.44	0.76	−1	1
sin(*Aspect*)	0.00	0.48	−1	1
cos(*Month*)	−0.60	0.56	−1	1
sin(*Month*)	0.03	0.56	−1	1
Wilderness	0.34	0.47	0	1
natural cause	0.56	0.50	0	1
year 2004	0.100	0.301	0	1
year 2005	0.203	0.403	0	1
year 2006	0.284	0.451	0	1
year 2007	0.184	0.388	0	1
year 2008	0.112	0.315	0	1
year 2009	0.105	0.306	0	1
region 2	0.056	0.230	0	1
region 3	0.083	0.276	0	1
region 4	0.353	0.478	0	1
region 5	0.069	0.253	0	1
region 6	0.107	0.309	0	1
region 8	0.182	0.386	0	1
region 9	0.015	0.122	0	1
region 10	0.070	0.256	0	1
fuel 1	0.121	0.327	0	1
fuel 2	0.066	0.247	0	1
fuel 3	0.024	0.152	0	1
fuel 5	0.119	0.324	0	1

a sample by truncation based on acreage is effectively imposing sample selection on costs per acre as well.

To examine the value of accounting for the bivariate relationship between cost per acre and truncated acreage, we estimate a univariate regression of cost without accounting for acreage, and compare the root mean squared forecast errors from this regression to that from the bivariate model.[12] The RMSE for the cost per acre equation from the bivariate model is 1.281, whereas the RMSE from the univariate regression is 1.513, suggesting that there is a loss of information and lower estimation precision with the univariate approach, because it does not account for correlation in the disturbances between cost per acre and acreage and therefore also does not incorporate the effects of acreage truncation on costs per acre.

Forecasts

We now provide a summary of in-sample predictions based on these regressions, the calculation of which is the primary objective of the development of this model. Table 4 and Fig. 1 provides summary statistics and distributions of both the unconditional predictions and the predictions conditional on the fire being selected into the sample based on fire size.

[12] The root mean squared error for regression equation j is calculated as $\text{RMSE}^j = \sqrt{\hat{\varepsilon}_c^{j\prime}\hat{\varepsilon}_c^{j}/(N-K)}$, where $\hat{\varepsilon}_c^{j}$ are the estimated errors from regression j.

Table 3
regression results.

Regressors	Dep. Vars.			
	ln(acres) [$\boldsymbol{a}$]		ln(cost/acre) [$\boldsymbol{c}$]	
	$\hat{\beta}$	$se(\hat{\beta})$	$\hat{\beta}$	$se(\hat{\beta})$
ln(*elevation*)	−0.0456	0.3143	0.2714	0.1523
latitude	−0.0975	0.0464	0.0016	0.0208
ln(*housing* 20)	−0.012	0.012	0.0127	0.0056
Energy Release	0.0326	0.0069	−0.0041	0.003
ln(*distance*)	0.5651	0.1536	−0.1831	0.0677
ln(*slope*)	0.2889	0.2482	0.4326	0.1266
cos(*Aspect*)	0.0691	0.1507	−0.1972	0.0708
sin(*Aspect*)	0.1547	0.1948	−0.1479	0.0954
cos(*Month*)	−0.6925	0.3442	−0.3782	0.1406
sin(*Month*)	0.2409	0.2393	−0.0279	0.0998
Wilderness	2.3419	0.2663	−0.6086	0.1217
natural cause	1.2234	0.2917	−0.2263	0.1235
year 2003			base	
year 2004	−2.0556	1.0613	1.1486	0.5029
year 2005	−2.0801	1.0427	0.8268	0.4928
year 2006	−2.1214	1.042	0.7845	0.4943
year 2007	−1.9702	1.0629	0.6929	0.5001
year 2008	−3.0124	1.0829	1.3108	0.5091
year 2009	−2.385	1.0931	0.6925	0.5091
region 1			base	
region 2	−1.7047	0.7083	−0.0046	0.3253
region 3	−1.6314	0.8587	0.2227	0.3856
region 4	0.4153	0.5519	−0.0088	0.2397
region 5	0.1005	0.7405	0.0976	0.3451
region 6	0.4512	0.5387	−0.2014	0.2381
region 8	−2.5477	0.9816	−0.6861	0.4266
region 9	−5.2231	1.7605	0.389	0.6373
region 10	5.332	1.1993	−2.2759	0.562
fuel 1	−0.1058	0.337	0.4495	0.1624
fuel 2	0.9401	0.4328	0.1245	0.2146
fuel 3	0.5295	1.034	1.1079	0.3802
fuel 4			base	
fuel 5	−0.3044	0.4052	0.9272	0.1801
constant	4.5941	2.5449	4.3807	1.115
σ_a	2.5677	0.1114		
σ_c			1.6648	0.0411
ρ	−0.5114	0.0302		
χ^2	198.4		*p* value	<0.0000
N	2061			

Table 4
Summary statistics for actual and expected acreage (A), cost per acre (C), and expenditures ($A{\cdot}C$). $N = 2061$.

Variable	Mean	S.D.	Min	Max
A	10,262	45,449	300	1,000,000
$E[A\|A \geq 100]$	20,300	50,451	595	818,026
$E[A]$	14,157	46,565	0	787,472
C	205	673	1	19,535
$E[C\|A \geq 100]$	206	197	7	2389
$E[C]$	783	747	21	9037
$A{\cdot}C$	592,724	2,586,133	372	83,700,000
$E[A{\cdot}C\|A \geq 100]$	857,505	1,287,676	3974	15,400,000
$E[A{\cdot}C]$	483,447	1,020,253	9	13,800,000

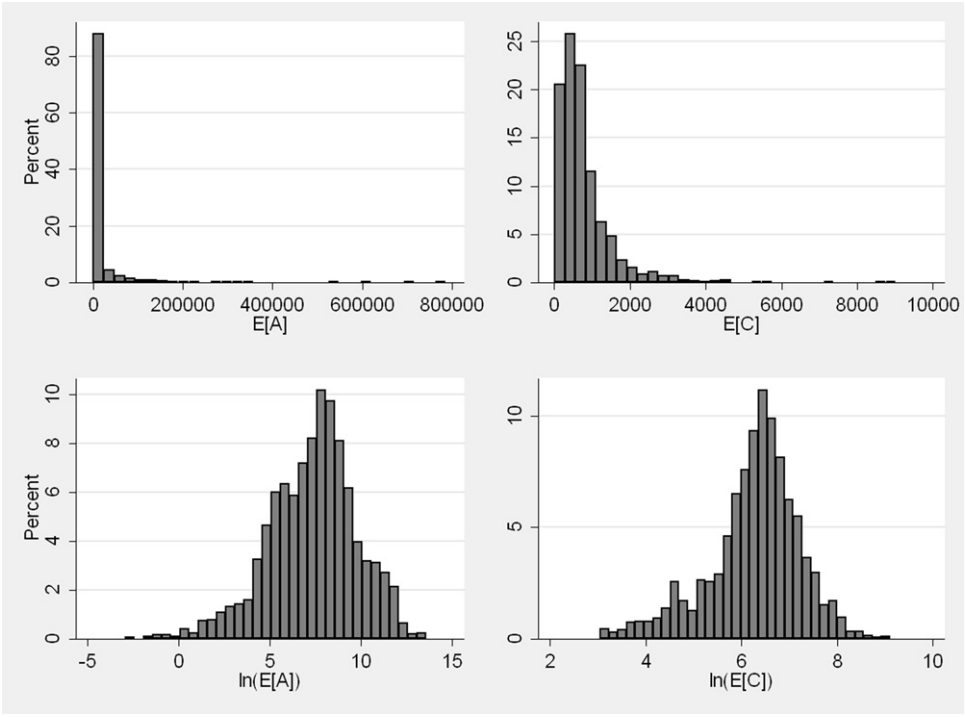

Fig. 1. Conditional and unconditional predicted values for acreage and cost/acre, and their logarithms.

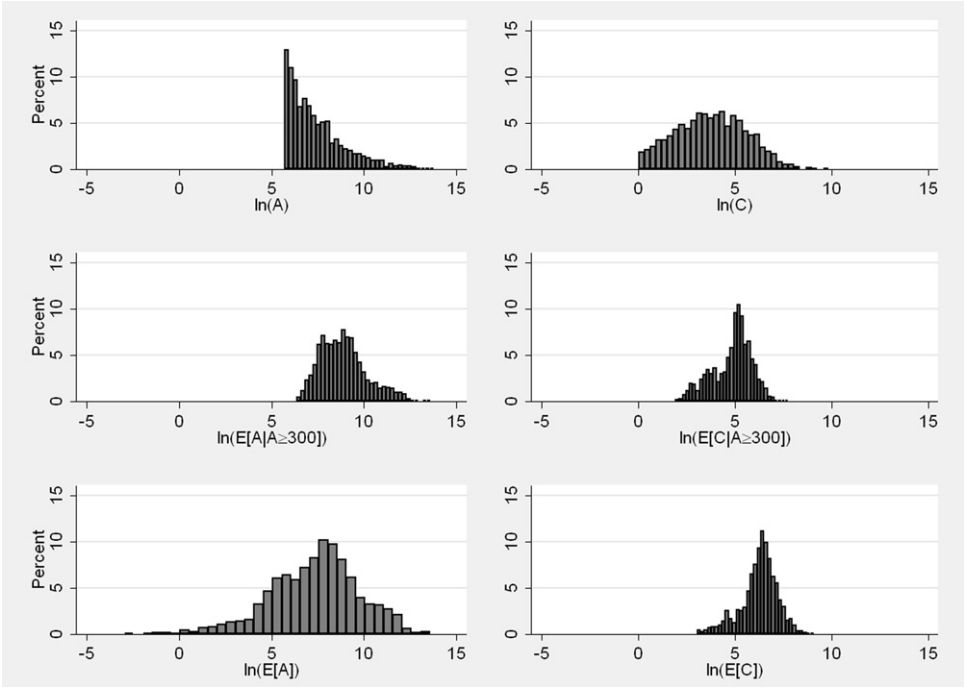

Fig. 2. Histograms of acreage and cost/acre data, and their conditional and unconditional predicated values.

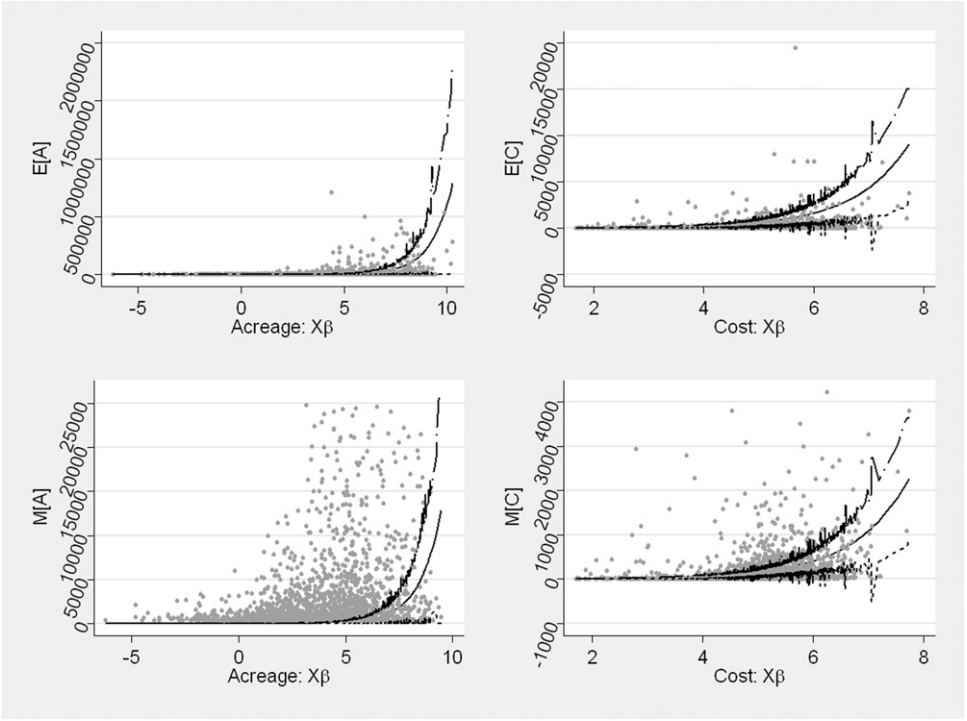

Fig. 3. Expected values (top row), and median values (bottom row), and 95% confidence bands for each. Actual values are gray dots (upper outliers truncated our of graph for illustrative purposes).

Fig. 1 shows expected acreage and costs, $E[A] \equiv E[Acres]$ and $E[C] \equiv E[Cost/Acre]$ in row 1, and their log transformations in row 2. The distributions of the expected values are highly skewed to the right, whereas the logarithms of these expectations are more symmetric.

Fig. 2 provides a comparison of the log transformation of the actual data, the expected acreage conditional on being greater than the minimum acreage of 300, and the expected acres for the population[s] as a whole. For the actual data, $\ln(acres) = a$ (column 1, row 1 in the figure), there are no observations below $\ln(300) = 5.7$, and there is a sharp truncation at this value, with several observations with A equaling exactly 300 acres (44 of 2061, or 2.1%). Note that all conditional predictions $\ln(E[A|A \geq 300])$ based on Eq. (17) fall at or above the minimum of $\ln(300) = 5.7$, as imposed by the model. However, the log of predicted acreage for the population as a whole, $E[A]$ (based on Eq. (11)), has a more symmetric distribution and ranges far below the sampling cutoff of $\ln(300) = 5.7$. This is precisely what we would expect, because the unconditional prediction pertains to the underlying population as a whole rather than only the subsample truncated at $A = 300$.

Fig. 2 column 2 shows the log transformations of actual cost per acre (row 1), predicted costs conditional on $A \geq 300$, and predicted costs regardless of whether $A \geq 300$. The distribution of the logarithm of conditional predictions $\ln(E[C|A \geq 300])$ tends to be farther to the left than the distribution of the log of unconditional predicted costs $\ln(E[C])$, which is loosely opposite of the case for the acreage predictions in column 1. This is because the correlation ρ between acreage and cost per acre is negative. Thus, the conditional predictions would tend to underestimate the costs per acre for the population as a whole because they pertain only to larger fires. Analogous to acreage, Eq. (12) should be used for estimating costs prior to knowing whether it will ultimately become 300 acres or more, and Eq. (18) should be used for fires of unknown final fire size but known to be larger than 300 acres.

Fig. 3 contains graphs of the unconditional expected values and median values for acreage and costs, as well as their estimated confidence intervals. For comparison, the actual values of acreage and

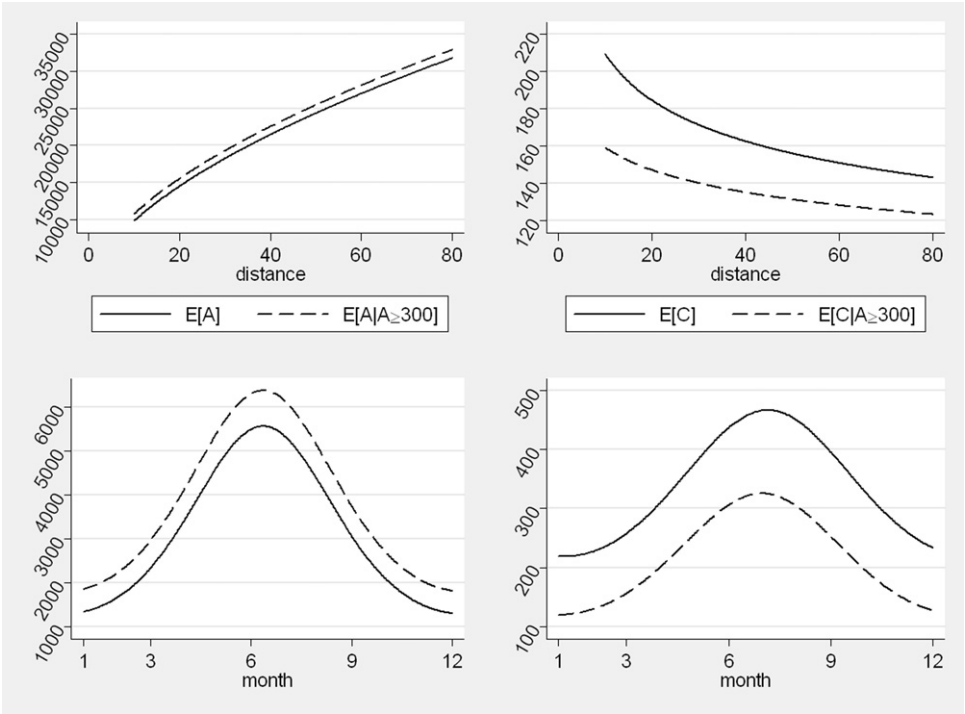

Fig. 4. Effect of month and distance on unconditional and conditional acreage and cost per acre predictions.

cost are included as gray scatter marks. However, note that these actual values are conditioned on being in the sample ($A \geq 300$). As such, the expected values and confidence intervals may appear "too low" for the acreage graphs, and too high for the cost graph relative to the actual plotted values.[13] The difference between the median and the expected values is $\exp(\sigma_z^2/2)$, for $Z \in (A, C)$. The expected value of fire size and acreage is much larger than the corresponding median values because of the skewness of the sampling distributions of these variables.

Fig. 4 allows a closer examination of the effects of individual variables on the conditional and unconditional predictions. In particular, consider the effects of the variable *distance* and *month* on the predicted acreage and costs. Fig. 4, row 1 column 1 shows the effect of *distance* on $E[A]$ and $E[A|A \geq 300]$. Row 1 column 2 shows the effect of *distance* on $E[C]$ and $E[C|A \geq 300]$ The prediction of $E[A|A \geq 300]$ tends to be higher for any given value of *distance* than does $E[A]$. In contrast, the effect of distance leads to a larger unconditional predicted C, but lower conditional predicted C. This difference is due again to a negative correlation coefficient ρ. The second row of Fig. 4 shows the effect of month on the conditional and unconditional predictions of A and C. Note that the peak fire size is slightly earlier in the year than the peak cost per acre, and that as before, conditional acreage is higher than unconditional acreage but the reverse is true for cost per acre.

Conclusion

This paper develops a method to generate *ex ante* predictions of fire size and costs per acre based on fire characteristics observable at the time of initial fire ignition, or at least prior to suppression

[13] It is also true that these 95% confidence intervals are likely to be too narrow even relative to estimation. The Delta method provides a linear approximation to the standard errors of the prediction that are likely to be an underestimate due to the nonlinearity of the exponential function used to calculate $E[A]$ and $E[C]$.

completion. As is common for wildfire modeling, our estimation efforts rely on a sample of data such that the fire records are systematically excluded from the sample if the final fire size is smaller than a lower bound. For data that are systematically sampled in this or similar fashion, the sampling process must be integrated into the modeling and estimation methods in order to generate statistically consistent parameter estimates and predictions. Further, fire acreage and cost/acre are often observed to be negatively correlated, and accounting for this correlation explicitly can help improve the efficiency (precision) of model estimates.

In order to account for these factors, we develop a Maximum Likelihood two-equation regression model with acreage-based truncation from below, assuming a bivariate lognormal disturbance process. We utilize the regression estimates to generate predicted values for acreage for the population as a whole, and for fires conditional on being a particular size or larger. The distributions of these model predictions allow distinctive illustrations of the importance of accounting for sample truncation when generating predicted outcomes based on *ex ante* information. The bivariate model that accounts for truncation appears to be promising for cost predictions as compared to a univariate model.

Acknowledgements

We thank two anonymous reviewers for insightful suggestions for this paper. This research received financial support from the U.S. Forest Service under Cooperative Agreement 09-CS-11221636-217 and the Washington State Agricultural Research Center under project #WPN00544.

References

Butry, D., Gumpertz, M., Genton, M., 2008. The production of large and small wildfires. In: Holmes, T.P., et al. (Eds.), The Economics of Forest Disturbances: Wildfires, Storms, and Invasive Species. , pp. 79–106 (Chapter 5).

Gebert, K., Calkin, D., Yoder, J., 2007. Estimating suppression expenditures for individual large wildland fires. Western Journal of Applied Forestry 22 (3), 188–196.

Greene, W., 2008. Econometric Analysis, sixth edition. Prentice Hall, Upper Saddle River, NJ.

Holmes, T., Huggett, R., Westerling, A., 2008. Statistical analysis of large wildfires. In: Holmes, T.P., et al. (Eds.), The Economics of Forest Disturbances: Wildfires, Storms, and Invasive Species. , pp. 59–77 (Chapter 4).

Kennedy, P.E., 1981. Estimation with correctly interpreted dummy variables in semilogarithmic equations. American Economic Review 71, 801.

Lien, D.-H.D., 1985. Moments of truncated bivariate log-normal distributions. Economics Letters 19, 243–247.

Lien, D., Balakrishnan, N., 2006. Moments and properties of multiplicatively constrained bivariate lognormal distribution with applications to futures hedging. Journal of statistical planning and inference 136, 1349–1359.

Mostafa, M.D., Mahmoud, M.W., 1964. On the problem of estimation for the bivariate lognormal distribution. Biometrika 51 (3/4), 522–527.

Strategic Issues Panel on Fire Suppression Costs, 2004. Large fire suppression costs: strategies for cost management. A report to the Wildland Fire Leadership Council. 59 pp.

U.S. Department of the Interior (USDI) and U.S. Department of Agriculture, Forest Service, 2011. FireCode System: Release 1.9. https://www.firecode.gov/help/User_Guide.pdf (accessed 17.10.11).

van Garderen, K.J., Shah, C., 2002. Exact interpretation of dummy variables in semilogarithmic equations. Econometrics Journal 5, 149–159.