THREE ESSAYS IN APPLIED LABOR AND RESOURCE ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Esther Washburn Mezey

August 2013

THREE ESSAYS IN APPLIED LABOR AND RESOURCE ECONOMICS

Esther Washburn Mezey, Ph.D.

Cornell University 2013

This dissertation includes three essays:

Spatio-temporal analysis of labor turnover statistics develops regularized dynamic spatio-temporal methods for selecting the sparse set of geographic, industrial, and demographic local-area job separation rates that best predict future rates in a given area using the public use Quarterly Workforce Indicators (QWI) made available by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program; it is the first paper to quantify the extent to which geographic, industrial, and demographic networks explain the dispersion of local-area labor market statistics through space and time.

Detecting and Repairing Data Errors in the Public Use Quarterly Workforce Indicator Data develops methods for detecting and repairing missing or erroneous observations in the QWI data.

Real Options in Resource Economics reviews the application of options valuation techniques from finance to capital budgeting decisions affecting real investments (real options) in the existing literature in forestry, fishery, water, and nonrenewable resources.

BIOGRAPHICAL SKETCH

Esther Washburn Mezey earned her Bachelor of Arts degree in Mathematics from Indiana University in 1996, after which she worked in the private sector in management consulting and software development. Her PhD research, completed at Cornell University, focuses on statistical and machine learning methods for the analysis of big data and spans three areas: forecasting local-area labor market statistics using sparse high-dimensional models (supervised learning, convex optimization, regularization, and principal component analysis), option valuation techniques applied to real investments in resource economics (stochastic calculus, dynamic programming, and contingent claims analysis), and outlier detection (unsupervised clustering). Her dissertation was supervised by Professors John M. Abowd and Jon M. Conrad.

ACKNOWLEDGEMENTS

I would like to thank John Abowd, Lars Vilhuber, and Erika McEntarfer for their supervision of my research using the public use Quarterly Workforce Indicators and gratefully acknowledge the support of the U.S. Census Bureaus's Longitudinal Employer-Household Dynamics (LEHD) Program, as well as the support of the National Science Foundation (through awards 9978093, 0339191, 0427889, 0922005, and 1131848) and the National Institute of Health (through NIA award 5R01AG0188854).

I would also like to thank Jon Conrad for his supervision of my research concerning real options in resource economics and acknowledge the support of the National Science Foundation through award 0832782.

TABLE OF CONTENTS

CHAPTER 1

SPATIO-TEMPORAL ANALYSIS OF LABOR TURNOVER STATISTICS

1.1 Introduction

When an economy contracts, how do job separation rates spread through the economy? In which industrial sectors and cities are quits, layoffs, and discharges first observed? Which sectors and cities follow? Why?

Motivated by these questions, we consider the high-dimensional problem of forecasting the one-period-ahead separations rate in a given local area based on separations rates in *all* local areas in the U.S. (which, collectively, we refer to as "neighborhood information"), and evaluate methods for reducing the dimensionality of the problem, including regularized approaches and methods using derived input directions. The aims of this study are twofold: first, we seek to determine whether models that include neighborhood information as inputs are superior to a naive model that excludes neighboring information and, second, we seek to establish data-driven methods for selecting relevant neighborhood information using state-of-the-art statistical and machine learning techniques.

1.2 Motivation and Literature Review

Increased interest in understanding the dynamics of local labor markets in the U.S. has led to the development of a new local-area statistical product, the Quarterly Workforce Indicators (QWI), produced by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program (Abowd et al., 2009;

Abowd and Vilhuber, 2011). The QWI provide quarterly local-area statistics on employment, accessions, separations, job creations, job destructions, and more, derived from the LEHD infrastructure files encompassing administrative data from State unemployment insurance (UI) systems. Unlike the Bureau of Labor Statistics' Local Area Unemployment Statistics (LAUS) program, which provides model-based estimates of monthly and annual employment, unemployment, and labor force measures, the QWI statistics are based almost entirely on employment counts, incorporating statistical methods only to protect the confidentiality of underlying entities in what is known as "disclosure proofing." The QWI, as a result, offer unprecedented detail on local labor market dynamics, making spatio-temporal analysis of these dynamics possible.

In this paper, we explore the notion that statistics in the QWI data that are near each other tend to be more alike than statistics that are far apart, where "near" refers to being proximal in terms of spatial networks and time, and consider a dynamic spatio-temporal model (DSTM)(Cressie and Wikle, 2011) where the statistic in a given location at time $t + 1$ depends on *all* local-area statistics in the data set at time t. That is, if y_t^i denotes the labor market statistic under consideration in local area i at time t, and there are a total of N local areas in the QWI data set, we consider a model of the form

$$y_{t+1}^i = f(\boldsymbol{x}_t, y_t^1, y_t^2, ..., y_t^N), \quad t = 1, ..., T$$

where $\boldsymbol{x}_t$ denote observable factors relating to all spatial locations, such as the national statistic and other macroeconomic factors.

The QWI provide statistics at the metropolitan, micropolitan, and county levels from approximately 2000 forward; thus, we have at our disposal N time series comprising $T \approx 44$ time periods, where $N \approx 300$ (if we consider spatio-

temporal dynamics among metropolitan areas only), $N \approx 800$ (if we consider spatio-temporal dynamics among both metropolitan and micropolitan areas), or $N \approx 3000$ locations (if we consider spatio-temporal dynamics among counties). With $N >> T$, we are faced with a high-dimensional problem where the number of unknown parameters exceeds the number of observations; as a result, we consider methods for "reducing the dimensionality" of the problem by searching for sparse solutions using a regularized likelihood approach or applying principal component or factor analysis.

Spatial and spatio-temporal analysis of local labor market statistics, such as labor force participation rates (Elhorst, 2008) and unemployment rates (Lopez-Bazo et al., 2005; Conley and Topa, 2002; Granger, 1969), can be found in the economics and regional sciences literature. High dimensional problems arise in macroeconomics (estimating multivariate or vector autoregressive models), forecasting (forecasting home prices using state-level or county-level data), and finance (estimating volatility matrices and forecasting financial returns); Fan et al. (2011) review the literature on sparse high-dimensional models and discuss applications in economics and finance; Bai and Ng (2008) and Stock and Watson (2006) provide a review of econometric analysis of dynamic factor models and forecasts based on principal components.

This paper is the first paper to analyze spatio-temporal patterns in local labor market statistics using statistical dimension reduction methods. Our results suggest new, improved methods for forecasting local-area statistics and constitute a significant contribution to the literature on spatio-temporal and high dimensional statistics, as well as the literature concerning labor and macroeconomics.

1.3 Data and Framework for Analysis

This study analyzes metropolitan- and micropolitan-area job separations-to-population rates (herein after referred to as separations rates) from 2000:Q3 to 2010:Q4 using QWI public-use data pertaining to the private sector. Separations rates are computed using the formula

$$y_t^i = \frac{s_t^i}{popest_{t-1}^i} * 100$$

where s_t^i denotes the number of separations in spatial location i during quarter t and $popest_{t-1}^i$ denotes the end-of-quarter population estimate in spatial location i in quarter $t-1$. Separations (s) are taken from the QWI pubic use data files; quarterly population estimates ($popest$) are derived from the annual core based statistical are population estimates available at http://www.census.gov/popest.

Metropolitan and micropolitan areas represent core urban areas in the U.S. with a population of 50,000 or more people (for metropolitan areas) or between 10,000 and 49,999 people (for micropolitan areas), and are defined by the Office of Management and Budget. We analyze the 805 metropolitan and micropolitan areas available in the QWI from 2000:Q3 to 2010:Q4, where each area conforms to the December 2009 Metropolitan and Micropolitan Statistic Area and Components definition file[1].

We employ a supervised learning approach that involves building a prediction model (learner) from a training set of data, selecting optimal tuning parameters for the prediction model, and assessing the prediction error of the optimized model over an independent set of test data. Accordingly, for each spatial area, we randomly select a block of 8 sequential observations as the indepen-

[1]See http://www.census.gov/population/metro/data/def.htm

dent test set (which we denote by *Test*) and use the remaining 30 observations as the training set (which we denote by *Train*). We carry out model selection on the training set and compute the mean squared prediction error over the test set to assess the generalization error of the models.

1.4 Model

We begin by considering a linear dynamical spatio-temporal model of the form

$$y_{t+1}^i = \theta_{i1} y_{t-3}^i + \theta_{i2} y_{t-4}^i + \theta_{i3} n_t + \theta_{i4} n_{t-3} + \theta_{i5} n_{t-4} + \sum_{j=1}^{n} \phi_{ij} y_t^j + \eta_{t+1}^i$$
$$(i = 1, ..., n; t = 1, ..., T), \tag{1}$$

where y_t^i denote the separations rate in spatial location i in period t, n_t denotes the nation-wide separations rate in period t, $\{\theta_{ik}\}_{k=1}^5$ and $\{\phi_{ij}\}_{j=1}^n$ are unknown parameters, and η_{t+1}^i is an unobservable error term assumed to be Gaussian.

For each spatial location i, this model comprises $5 + n$ unknown parameters (where $n = 805$, the number of metropolitan and micropolitan areas included in our sample) and T observations (where $T = 38$, the number of quarters in our sample), resulting in a "high-dimensional" problem. We seek to reduce the dimensionality of the problem by

1. eliminating the predictors $\{y_t^j\}_{j \neq i}$, generating a naive model,
2. searching for sparse solutions using penalized models, and
3. applying principal component analysis to summarize neighborhood information in a reduced set of variables.

The sections that follow describe each method in detail.

1.4.1 Naive Model

One way to significantly reduce the dimensionality of model (1) is to exclude "neighborhood information" and evaluate a model that relates the separations rate in spatial area i to its own lagged rates in addition to lagged nation-wide rates,

$$y^i_{t+1} = \phi_{ii} y^i_t + \theta_{i1} y^i_{t-3} + \theta_{i2} y^i_{t-4} + \theta_{i3} n_t + \theta_{i4} n_{t-3} + \theta_{i5} n_{t-4} + \eta^i_{t+1} \qquad (t = 1, ..., T). \quad (2)$$

This model serves as a baseline model with which to compare the penalized and principal components regression models described below, which include neighborhood information.

Estimation and Model Assessment

We use OLS to estimate model (2) from the training data and then compute the Ljung-Box test statistic of its residual series to verify that this series behaves as white noise. We find that, for 94 percent of the spatial areas included in our sample, the Ljung-Box test statistics has a p-value less than .05, indicating that the residual series for these areas are not serially correlated. Thus, we conclude that the naive model is a reasonable representation of the job separations process.

To assess prediction capability of the naive model, we compute the mean squared prediction error over the independent test sample,

$$MSPE^{i,naive}_{test} \equiv \frac{1}{V} \sum_{t \in Test} \left[y^i_{t+1} - \hat{y}^i_{t+1} \right]^2,$$

where V denotes the number of elements in the test set and $\hat{y}^i_{t+1}$ denotes the predicted value of y^i_{t+1} generated using parameters estimated from the training

set. That is, the parameter estimates are given by

$$\{\hat{\theta}_{ik}\}_{k=1}^{5}, \hat{\phi}_{ii} =$$
$$\arg\min_{\{\theta_{ik}\},\phi_{ii}} \sum_{s\in Train} \left(y_{s+1}^{i} - (\phi_{ii}y_{s}^{i} + \theta_{i1}y_{s-3}^{i} + \theta_{i2}y_{s-4}^{i} + \theta_{i3}n_{s} + \theta_{i4}n_{s-3} + \theta_{i5}n_{s-4})\right)^{2}.$$

1.4.2 Penalized Models: Lasso, Ridge, Elastic Net

For each spatial location i, model (1) is a candidate for the following regularized linear regression, which shrinks the coefficients $\{\phi_{ij}\}_{j\neq i}$ by imposing a penalty on their size,

$$\min_{\{\theta_{ik}\},\{\phi_{ij}\}} \sum_{t=1}^{T} \left(y_{t+1}^{i} - \theta_{i1}y_{t-3}^{i} - \theta_{i2}y_{t-4}^{i} - \theta_{i3}n_{t} - \theta_{i4}n_{t-3} - \theta_{i5}n_{t-4} - \sum_{j=1}^{n}\phi_{ij}y_{t}^{j}\right)^{2}$$
$$\text{subject to } \sum_{j\neq i}\omega_{ij}^{(p)}\left[\frac{(1-\alpha)}{2}\phi_{ij}^{2} + \alpha|\phi_{ij}|\right] < c. \tag{3a}$$

In this formulation, $c \geq 0$ is a user-defined parameter that controls the amount of shrinkage applied to the estimates $\{\phi_{ij}\}_{j\neq i}$ and

$$P_{\alpha,p}(\{\phi_{ij}\}_{j\neq i}) = \sum_{j\neq i}\omega_{ij}^{(p)}\left[\frac{(1-\alpha)}{2}\phi_{ij}^{2} + \alpha|\phi_{ij}|\right]$$

is a penalty function that accommodates the lasso ($\alpha = 1$), the ridge ($\alpha = 0$) and the elastic net [$\alpha \in (0, 1)$] and permits the specification of different penalties for the coefficients $\{\phi_{ij}\}_{j\neq i}$ through the penalty scaling parameters $\{\omega_{ij}^{(p)}\}_{j\neq i}$.

The optimization problem in (3a) can be equivalently expressed in its Lagrangian (or penalized) form

$$\min_{\{\theta_{ik}\},\{\phi_{ij}\}} \sum_{t=1}^{T} \left(y_{t+1}^{i} - \theta_{i1}y_{t-3}^{i} - \theta_{i2}y_{t-4}^{i} - \theta_{i3}n_{t} - \theta_{i4}n_{t-3} - \theta_{i5}n_{t-4} - \sum_{j=1}^{n}\phi_{ij}y_{t}^{j}\right)^{2}$$
$$+\lambda P_{\alpha,p}(\{\phi_{ij}\}_{j\neq i}), \tag{3b}$$

where $\lambda \geq 0$ is a user-defined tuning parameter that controls the amount of shrinkage applied to the coefficients $\{\phi_{ij}\}_{j \neq i}$ (there is a one-to-one correspondence between λ and c).

Penalty scaling parameters are chosen according to the distances between spatial locations i and j, $j = 1, ..., n$. We consider several distance metrics in our analysis, including

- population-weighted geographic distance ($p = 6$), which measures the population-weighted great-circle distance between two spatial locations, and
- population-weighted industrial distance ($p = 7$), which measures the population-weighted industrial distance between two spatial locations, where industrial distance is measured by the euclidean distance between the vector of percentages of wages paid to workers in different NAICS sectors, and
- "no distance weights" ($p = 5$), which assumes $\omega_{ij} = 1$ for all i, j.

If $\omega_{ik}^{(p)} < \omega_{ij}^{(p)}$, spatial location k is "nearer" to location i than location j; as a result, coefficient k is penalized less than coefficient j (less shrinkage is applied to coefficient k), increasing the likelihood that the estimate of coefficient k is nonzero and the estimate of coefficient j is zero (or arbitrarily close to zero).

Why include penalty scaling parameters? The separations rate time series in our model are highly correlated. As a result, the inclusion of penalty scaling parameters may result in more economically meaningful results, where the spatial locations associated with the non-zero coefficients obtained in the penalized regression may represent "predictive" or "causal" factors.

For more information on these penalty scaling parameters, refer to the data appendix.

Estimation

Given values of α_i and p_i, $\alpha_i \in \{0, 0.25, 0.50, 0.75, 1\}$ and $p_i \in \{5, 6, 7\}$, we use the *glmnet* R package (Friedman et al., 2010; Simon et al., 2011) to generate a grid of values for λ_i and estimate the coefficients of the penalized model for each value of λ_i.

Model Selection

The penalized regression in equation (3) includes three tuning parameters

- α_i, which determines the elastic net mix,
- p_i, which determines the distance metric on which the penalty scaling parameters are based, and
- λ_i, which controls the amount of shrinkage applied to the coefficients,

producing a set of models indexed by triples (α_i, p_i, and λ_i).

We estimate the prediction error of these models using leave-one-out cross-validation, a method that exploits efficient reuse of the training data, and then choose the values of α_i, p_i, and λ_i that minimize this error.

Leave-one-out Cross-Validation

Given values of α_i and p_i, we use the *glmnet* R package (Friedman et al., 2010; Simon et al., 2011) to generate a grid of values for λ_i, and then compute leave-one-out cross validation error using the *cv* function in the *glmnet* package. Once we have computed $CV(\alpha_i, p_i, \lambda_i)$ for all triples $(\alpha_i, p_i, \lambda_i)$, we select the values of α_i, p_i, and λ_i that minimize $CV(\alpha_i, p_i, \lambda_i)$, denoting these optimal values by $\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i$.

Model Selection Results

Table 1.1 lists the median number of nonzero coefficients obtained for different values of α and p (and the value of λ that minimizes cross-validation error for these values of α and p). Here we observed that the number of nonzero coefficients increases as α decreases. This is as expected: elastic net regularization is a convex combinations of the ridge ($\alpha = 0$) and the elastic net ($\alpha = 1$); as such, it behaves increasingly like the ridge regression as α tends toward zero. We also observe that the use of population-weighted geographic penalty scaling parameters leads to a more complex model (with more nonzero coefficients) than those obtained with the population-weighted industrial penalty scaling parameters. This suggests that, on average, spatial areas tend to have more "close" geographic neighbors than industrial neighbors.

Leave-one-out cross-validation favors the lasso (in 35 percent of spatial areas), followed by the ridge (in 34 percent of spatial areas), followed by the elastic net (in 31 percent of spatial areas). In terms of penalty scaling parameters, it favors the use of no weights (in 59 percent of spatial areas), followed by geographic penalty scaling parameters (in 23 percent of spatial areas), followed by industrial penalty scaling parameters (in 19 percent of spatial areas).

	Penalty Scaling Parameters		
Method	None ($p = 5$)	Pop-weighted geographic ($p = 6$)	Pop-weighted industrial ($p = 7$)
Lasso ($\alpha = 1$)	13	13	11
Elastic net ($\alpha = 0.75$)	14	13	12
Elastic net ($\alpha = 0.5$)	17	16	15
Elastic net ($\alpha = 0.25$)	24	23	20
Ridge ($\alpha = 0$)	810	810	810

Table 1.1: Median number of nonzero coefficients obtained when minimizing CV error

Figure 1.1 portrays the optimal penalty scaling parameters grouped by the population size of the spatial areas. As the population size increases, "no weights" becomes increasing optimal and the population-weighted geographic and industrial weights become decreasingly optimal. This suggests that geographic and industrial networks may be more important in explaining labor turnover in small areas.

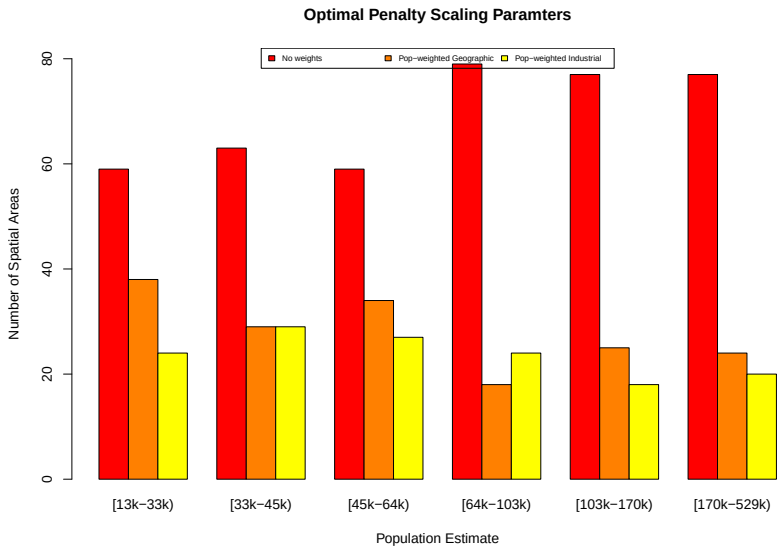


Figure 1.1: Optimal penalty scaling parameters grouped by population size

Model Assessment

To assess the prediction capability of the model with optimal tuning parameters $\hat{\alpha}_i$, $\hat{p}_i$, and $\hat{\lambda}_i$, we compute the mean squared prediction error over the independent test sample,

$$MSPE^{i,penalized}_{test} \equiv \frac{1}{V} \sum_{t \in Test} \left[y^i_{t+1} - \hat{y}^i_{t+1|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i} \right]^2,$$

where V denotes the number of elements in the test set and $\hat{y}^i_{t+1|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i}$ denotes the predicted value of y^i_{t+1} generated using the parameters $\{\hat{\theta}_{ik|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i}\}_{k=1}^5$ and $\{\hat{\phi}_{ij|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i}\}_{j=1}^n$ estimated from the training set. That is,

$$\{\hat{\theta}_{ik|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i}\}_{k=1}^5, \{\hat{\phi}_{ij|\hat{\alpha}_i, \hat{p}_i, \hat{\lambda}_i}\}_{j=1}^n =$$

$$\arg \min_{\{\theta_{ik}\}, \{\phi_{ij}\}} \sum_{s \in Train} \left(y^i_{s+1} - \theta_{i1} y^i_{s-3} - \theta_{i2} y^i_{s-4} - \theta_{i3} n_s - \theta_{i4} n_{s-3} - \theta_{i5} n_{s-4} - \sum_{j=1}^{n} \phi_{ij} y^j_s \right)^2$$

$$+ \hat{\lambda} P_{\hat{\alpha}, \hat{p}}(\{\phi_{ij}\}_{j \neq i})$$

where $\hat{\alpha}_i$,$\hat{p}_i$, and $\hat{\lambda}_i$ are the optimal tuning parameters that minimize leave-one-out cross-validation error.

1.4.3 Principal Component Regression

Model (1) is also a candidate for a principal component regression, incorporating factors $\widehat{F^i_{mt}}, m = 1, ..., M_i$, that are linear combinations of the original $N - 1$ inputs $\{y^j_t\}_{j \neq i}$, estimated by the method of principal components.[2] Principal component analysis takes a linear transformation of the data $\{y^j_t\}_{j \neq i}$ to a new coordinate system such that the greatest variance by any projection of the data lies

[2] A principal component regression is referred to as a dynamic factor model, a factor-augmented regression, or factor-augmented vector autoregression (FAVAR) in the time series literature.

on the first coordinate (called the first principal component), the second greatest variance lies on the second coordinate (called the second principal component), and so on, and the covariance between each principal component is zero; in essence, principal component analysis provides a lower-dimensional "picture" of the data $\{y_t^j\}_{j\neq i}$ that captures as much information as possible from the higher-dimensional space.

We evaluate the following principal component regression,

$$y_{t+1}^i = \phi_{ii} y_t^i + \theta_{i1} y_{t-3}^i + \theta_{i2} y_{t-4}^i + \theta_{i3} n_t + \theta_{i4} n_{t-3} + \theta_{i5} n_{t-4} + \sum_{m=1}^{M} \psi_{im} \widehat{F}_{mt}^i + \eta_{t+1}^i \qquad (t = 1, ..., T), \tag{4}$$

where y_s^i, n_s^i, $s = t, t-3, t-4$, and the estimated factors $\{\widehat{F}_{mt}^i\}$ are treated as observable inputs, $\{\theta_{ik}\}$, ϕ_{ii}, and $\{\psi_{im}\}$ are unknown parameters, and η_{t+1}^i is an unobservable error term that is assumed to be Gaussian.

Estimation

Equation (4) is estimated in two steps. First, we use principal component analysis to estimate the factors $\widehat{F}_{mt}^i$, $m = 1, ..., M, t \in Train$, following these steps:

1. Form the U x $(N-1)$ matrix of separations rate data, $\{y_t^j\}_{j\neq i, t\in Train}$, where U is the number of elements in the training sample, and denote it by $\boldsymbol{y}_{Train}$.
2. Run the *prcomp* R package Mardia et al. (1979); Venables and Ripley (2002) to perform principal component analysis on the covariance matrix of $\boldsymbol{y}_{Train}$, $\boldsymbol{\Sigma}_{y_{Train}}$, obtaining the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}_{y_{Train}}$, $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_U, \boldsymbol{e}_U)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_U \geq 0$ and $(\lambda_t, \boldsymbol{e}_t)$ is associated

with component t. Compute the U-dimensional factors (also called scores) $\widehat{\boldsymbol{F}}_m^i, m = 1, ..., M$, as

$$\widehat{\boldsymbol{F}}_m^i = \boldsymbol{y}_{Train}\boldsymbol{e}_m.$$

For a given spatial location i, we then use OLS to estimate the unknown parameters $\{\theta_{ik}\}$, ϕ_{ii}, and $\{\psi_{im}\}$.

Model Selection

The principal component regression of equation (4) includes the tuning parameter M_i, which determines the number of principal components in the model, producing a set of models indexed by M_i. We estimate the prediction error of these models using leave-one-out cross-validation error, $CV(M_i)$. Once we have computed $CV(M_i)$ for all candidate values of M_i, we select the value of M_i that minimizes $CV(M_i)$ and denote it by $\widehat{M}_i$.

Model Selection Results

The exercise of minimizing cross-validation error $CV(M_i)$ suggests that including one principal component (in 368 spatial locations), two principal components (in 124 locations), three components (in 90 locations), or four or more components (in 223 location) minimizes the prediction error of these models.

Table 1.2 presents the results of principal component analysis for the first six principal components generated from the separations rate data excluding area 10020 (Abbeville, LA); the first component accounts for 55 percent of the total variability of the data and the first three components, together, account for 76

percent of the total variability.

Component	1	2	3	4	5	6
Eigenvalue	1648	424	201	129	92	74
Proportion of total variance	.553	.142	.067	.043	.031	.025
Cumulative variance	.553	.695	.762	.805	.836	.861

Table 1.2: Results of principal component analysis for the $N-1$ metropolitan-area separations rates (excluding Abbeville LA), 2000:Q3 to 2008:Q3 (first six components)

Figure 1.2 presents a scree plot (a plot of eigenvalues ordered from largest to smallest) for the $N-1$ metropolitan-area separations rates excluding area 10021 (Abbeville, LA). Tsay Tsay (2010) recommends looking for an "elbow" in the scree plot, indicating that the remaining eigenvalues are relatively small and about the same size, as an informal but effective method for determining the number of components to include in an application. This method suggests that the first two or three components explain most of the variation in the data, and confirms the model selection results obtained through cross-validation error minimization.

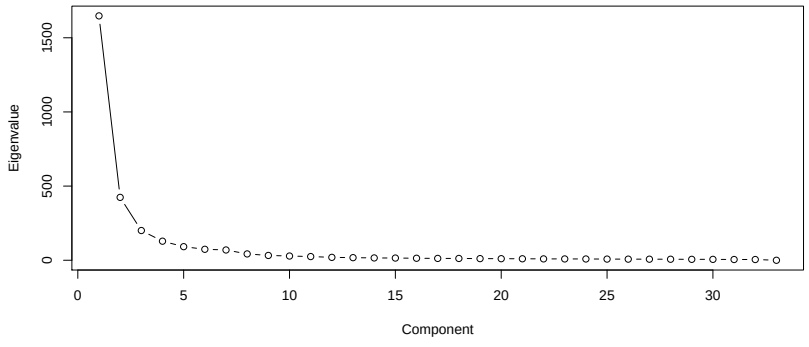


Figure 1.2: Scree plot for the $N-1$ metropolitan-area separations rates (excluding Abbeville LA), 2000:Q3 to 2008:Q3

Model Assessment

To assess the prediction capability of the principal component regression with optimal tuning parameter $\widehat{M}_i$, we compute the mean squared prediction error over the independent test sample,

$$MSPE^{i,pc}_{test} \equiv \frac{1}{V}\sum_{t\in Test}\left[y^i_{t+1} - \hat{y}^i_{t+1|\widehat{M}_i}\right]^2,$$

where $\hat{y}^i_{t+1|\widehat{M}_i}$ denotes the predicted value of y^i_{t+1} generated using parameter estimates $\{\hat{\theta}_{ik|\widehat{M}_i}\}$, $\hat{\phi}_{ii|\widehat{M}_i}$, and $\{\hat{\psi}_{im|\widehat{M}_i}\}_{m=1}^{\widehat{M}_i}$ given by

$$\{\hat{\theta}_{ik|\widehat{M}_i}\}, \hat{\phi}_{ii|\widehat{M}_i}, \{\hat{\psi}_{im|\widehat{M}_i}\}_{m=1}^{\widehat{M}_i} =$$

$$\arg\min_{\{\theta_{ik}\},\phi_{ii}\{\psi_{im}\}} \sum_{s\in Test}\left(y^i_{s+1} - \phi_{ii}y^i_s - \theta_{i1}y^i_{s-3} - \theta_{i2}y^i_{s-4} - \theta_{i3}n_s - \theta_{i4}n_{s-3} - \theta_{i5}n_{s-4} - \sum_{m=1}^{\widehat{M}_i}\psi_{im}\widehat{F}^i_{ms}\right)^2.$$

It is assumed that $\widehat{F}^i_{ms}$, $s \in Test$ are estimated following these steps:

1. Form the V x $(N-1)$ matrix of separations rate data, $\{y^j_s\}_{j\neq i, s\in Test}$, where V is the number of elements in the test sample, and denote it by $\boldsymbol{y}_{Test}$.
2. Compute the V-dimensional factors (also called scores) $\widehat{\boldsymbol{F}}^i_m, m = 1, ..., \widehat{M}$, as

   $$\widehat{\boldsymbol{F}}^i_m = \boldsymbol{y}_{Test}\boldsymbol{e}_m.$$

   where $\boldsymbol{e}_m$, $m = 1, ..., \widehat{M}$, are the eigen vectors obtained from the eigen decomposition of the covariance matrix of $\boldsymbol{y}_{Train}$.

1.5 Results

The prediction capability of the naive, principal component regression, and penalized models is assessed by comparing the mean squared prediction error

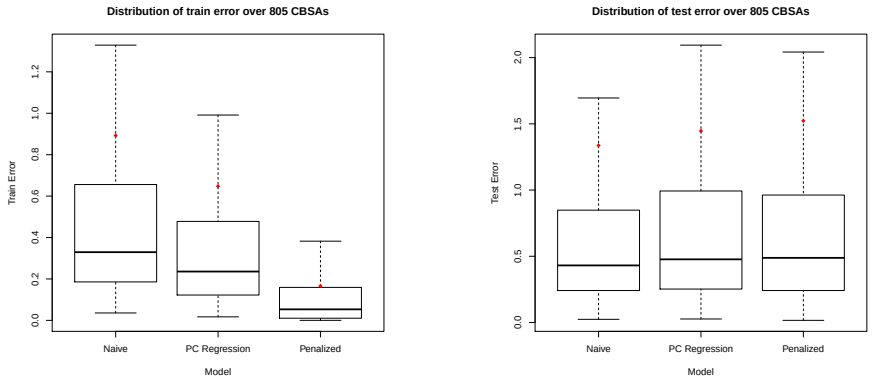
(MSPE) over the training and independent test samples for each spatial area included in the analysis. Table 1.3 (1.4) presents summary statistics that describe the distribution of the training (test) MSPE across spatial areas, and figure 1.3 presents box plots of these distributions.

Model	Mean	Median	Median Absolute Deviation about the Median
Naive	.89	.33	.18
PC Regression	.65	.24	.13
Penalized	.17	.05	.05

Table 1.3: Summary Statistics: Distribution of the Training MSPE across the 805 Spatial Areas included in our Analysis

Model	Mean	Median	Median Absolute Deviation about the Median
Naive	1.34	.43	.25
PC Regression	1.45	.48	.29
Penalized	1.52	.49	.30

Table 1.4: Summary Statistics: Distribution of the Test MSPE across the 805 CBSAs included in our Analysis



(a) (b)

Figure 1.3: Distribution of the Training and Test MSPE

The penalized model outperforms the naive and principal component re-

gression models in the training set, but underperforms these models in the independent test set. One explanation for the poor performance in the independent test sample is the size of our training and test samples: with 30 observations, the training sample may not be sufficiently large to generate meaningful parameter estimates, and with 8 observations, the test sample may not be sufficiently large to adequately demonstrate the predictive capability of the penalized model. Another explanation for the poor performance is the high multicollinearity seen in the data. Future research should examine the fused lasso and weight-fused elastic net, which incorporate different "grouping" mechanisms for selecting predictors when the data exhibit high multicollinearity and which permit the specification of weights based on correlation and other user-defined parameters.

1.6 Appendix A: The Data

This study uses 2000:Q3 to 2011:Q1 metropolitan- and micropolitan-area QWI public use data pertaining to the private sector (QWI ownership codes *A*05) available for download at http://www.vrdc.cornell.edu/qwipu/.

The following data files are used:

- "qwi_us_wia_county_naicssec_pri" (age group x gender tabulations)
- "qwi_us_wia_metro_naicssec_pri" (age group x gender tabulations)
- "qwi_us_rh_metro_naicssec_pri" (ethnicity and race tabulations)

1.6.1 Separations to Population Rates

Separations to population rates are constructed using the formula

$$y_t^i = \frac{s_t^i}{popest_{t-1}^i} * 100$$

where s_t^i denotes the number of separations in spatial location i during quarter t and $popest_{t-1}^i$ denotes the end-of-quarter population estimate in spatial location i in quarter $t-1$.

Separations (s) are taken from the QWI pubic use data files.

Quarterly population estimates (*popest*) are derived from the annual CBSA population estimates available at http://www.census.gov/popest.

We first construct series of annual, county-level population estimates spanning 2000 to 2011 using the Vintage 2011 County Population dataset (with estimates for 2010 and 2011) and the County Intercensal Estimates (with estimates from 2000 to 2010). Intercensal estimates reconcile postcensal estimates (from 2000 to 2010) "with the 2010 Census counts and provide a consistent time series of population estimates that reflect the most recent census results."[3]

These annual estimates represent the number of people living in a county as of a specified point in time (usually July 1st) in a given year. Accordingly, we assume annual estimates represent the county population at the end of the second quarter and use linear interpolation to derive population estimates for the end of the first, third, and fourth quarters within the series.

We then summarize the quarterly, county-level estimates at the CBSA-level

[3]See the Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010 at http://www.census.gov/popest/methodology/2000-2010_Intercensal_Estimates_Methodology.pdf

using the December 2009 metropolitan and micropolitan statistical areas and components definition file available at:

http://www.census.gov/population/metro/data/def.html.

1.6.2 Penalty Scaling Parameters (used in penalized regression)

In the regularization methods used in this paper, we allow the specification of different penalties for each of the coefficients using penalty scaling parameters $\omega_{ij} \geq 0, j = 1, ..., n$. If $\omega_{ij} > 0$, then the penalty applied to ϕ_{ij} is $\omega_{ij}\lambda$; if $\omega_{ij} = 0$, the coefficient does not get penalized. In general, if $\omega_{ij} > \omega_{ik}$ then coefficient j is penalized more than coefficient k (more shrinkage is applied to the coefficient j), increasing the likelihood that coefficient j is zero and coefficient k is nonzero.

Different penalty scaling parameters are described in the sections that follow.

Geographic Distance Penalty Scaling Parameters

The geographic distance between CBSA i and j, GD_{ij}, is computed using the haversine formula,

$$GD_{ij} = 2r \arcsin\left(\sqrt{\sin^2\left((lt_i - lt_j)/2\right) + \cos(lt_i)\cos(lt_j)\sin^2\left((ln_i - ln_j)/2\right)}\right),$$

where r is the radius of the earth (3963.1676 miles), and lt_k and ln_k are the latitude and longitude of the internal point (centroid) of CBSA k, $k = i, j$.

The penalty scaling parameters that correspond to geographic distances are computed as

$$\omega_{ij} = \frac{GD_{ij}}{\sum_{k=1}^{n} GD_{ik}}$$

Population-Weighted Geographic Distance Penalty Scaling Parameters

The population-weighted geographic distance between CBSA i and j, $PWGD_{ij}$, is computed as

$$PWGD_{ij} = \frac{\frac{1}{\widetilde{pop}^{j}_{2000}} * \widetilde{GD}_{ij} + \left| \min\left\{ \frac{1}{\widetilde{pop}^{k}_{2000}} * \widetilde{GD}_{ik} \right\}_{k=1}^{n} \right| + \epsilon}{\sum_{k=1}^{n} \left(\frac{1}{\widetilde{pop}^{k}_{2000}} * \widetilde{GD}_{ik} + \left| \min\left\{ \frac{1}{\widetilde{pop}^{k}_{2000}} * \widetilde{GD}_{ik} \right\}_{k=1}^{n} \right| + \epsilon \right)}$$

where $\widetilde{pop}^{j}_{2000}$ is the standardized population estimate of spatial area j in 2000, and $\widetilde{GD}_{ij}$ is the standardized geographic distance between CBSA i and j, GD_{ij}, and $\epsilon > 0$ is a constant. The geographic distance, GD_{ij}, is computed using the haversine formula,

$$GD_{ij} = 2r \arcsin\left(\sqrt{\sin^2\left((lt_i - lt_j)/2\right) + \cos(lt_i)\cos(lt_j)\sin^2\left((ln_i - ln_j)/2\right)} \right),$$

where r is the radius of the earth (3963.1676 miles), and lt_k and ln_k are the latitude and longitude of the internal point (centroid) of CBSA k, $k = i, j$.

The penalty scaling parameters that correspond to geographic distances are computed as

$$\omega_{ij} = (PWGD_{ij})^{\gamma}.$$

where $\gamma = 3$ is a tuning parameter. These parameters are

- increasing in distance (higher penalties are applied to spatial areas that are farther away),
- decreasing in the population (smaller penalties are applied to areas with larger populations),
- standardized (so that population and distance carry equal weight in the final parameters), and

- strictly positive (penalty scaling parameters must be strictly greater than zero; if a penalty scaling parameter is negative,the associated coefficient is amplified, and if a penalty scaling parameter is zero, the associated coefficient is not penalized).

Industrial Distance Penalty Scaling Parameters

Using an approach similar to that employed by Conley and Topa Conley and Topa (2002), the industrial distance between CBSA i and j, ID_{ij}, is computed as the Euclidian distance between the vector $\boldsymbol{IS}_i$ of percentages of wages paid to workers in different NAICS sectors in CBSA i and the corresponding vector $\boldsymbol{IS}_j$ in CBSA j.

$$ID_{ij} = \sqrt{\sum_{k=1}^{19}(IS_{ik} - IS_{jk})^2}$$

Table 4 presents a list of NAICS sectors included in our analysis (sector 92, public administration, is not included, since our analysis concerns the private sector only). Two CBSAs with the same industrial composition will have an industrial distance of zero.

The penalty scaling parameters that correspond to industrial distances are computed as

$$\omega_{ij} = ID_{ij}, j = 1, ..., n$$

Ethnic and Racial Penalty Scaling Parameters

Using an approach similar to that employed by Conley and Topa Conley and Topa (2002), the ethnic and racial distance between CBSA i and j, ED_{ij}, is com-

puted as the Euclidian distance between the vector $\boldsymbol{ER}_i$ of percentages of workers who are black, white, asian, and hispanic in CBSA i and the corresponding vector $\boldsymbol{ER}_j$ in CBSA j.

$$ED_{ij} = \sqrt{\sum_{k=1}^{4}(ER_{ik} - ER_{jk})^2}$$

The penalty scaling parameters that correspond to ethnic and racial distances are computed as

$$\omega_{ij} = ED_{ij}, j = 1, ..., n$$

1.7 Appendix B: Spatio-Temporal Analysis of Unemployment Rates Using LAUS Data

The QWI data are available from 2000, limiting our analysis sample to ~40 time periods. To get a sense of how the penalized models presented in this paper perform over a longer time frame, we analyze metropolitan-area unemployment rates in the LAUS data, which are available on a monthly basis from 1990, yielding a data set of ~250 time periods.

1.7.1 Introduction

We hypothesize that the change in the unemployment rate in period $t + 1$ in a given metropolitan area depends on changes in the unemployment rate in other metropolitan areas in period t (such as neighboring areas or areas with close industrial ties to the response area) but we don't know which neighbors in period t are relevant. Thus we consider the high-dimensional problem of

relating future changes in the unemployment rate in a given metropolitan area to changes in all 380 metropolitan areas in the U.S. and evaluate methods for reducing the dimensionality of the problem.

1.7.2 Framework for Analysis

We employ a supervised learning approach that involves building a prediction model (learner) from a training set of data, selecting optimal tuning parameters for the prediction model, and assessing the prediction error of the optimized model over an independent set of test data. According, we divide our data into a training sample (with 177 observations from April 1990-December 2004) and a test sample (with 82 observations from January 2005 - November 2011), and carry out model selection on the training sample using cross-validation.

1.7.3 Naive Model

Letting y_t^i denote the change in the unemployment rate in spatial location i in period t and letting $\boldsymbol{x}_t$ denote a vector containing seasonal dummy variables and the change in the national unemployment rate in period t, the naive model is given by

$$y_{t+1}^i = \boldsymbol{x}_t'\boldsymbol{\theta}_i + \sum_{l=0}^{L-1} \pi_{il} y_{t-l}^i + \eta_{t+1}^i, \quad (i = 1, ..., N; t = L, ..., T)$$

where $\boldsymbol{\theta}_i$ and $\pi_{il}, l = 0, ..., L-1$ are unknown parameters and η_{t+1}^i is an unobservable error term assumed to be independent over time and Gaussian.

1.7.4 Penalized Model

The penalize model is given by

$$\min_{\boldsymbol{\theta_i},\{\pi_{il}\},\{\phi_{ij}\}} \quad \sum_{t=L}^{T}\Big(y_{t+1}^{i} - \boldsymbol{x}_t'\boldsymbol{\theta}_i - \sum_{l=0}^{L-1}\pi_{il}y_{t-l}^{i} - \sum_{j\neq i}\phi_{ij}y_t^{j}\Big)^2 \; + \lambda P_{\alpha,p}\Big(\{\phi_{ij}\}_{j\neq i}\Big)$$

where λ is a tuning parameter that controls the amount of shrinkage applied to the coefficients $\{\phi_{ij}\}_{j\neq i}$, and

$$P_{\alpha,p}\Big(\{\phi_{ij}\}_{j\neq i}\Big) = \sum_{j\neq i}\omega_{ij}^{(p)}\left[\frac{(1-\alpha)}{2}\phi_{ij}^2 + \alpha|\phi_{ij}|\right]$$

is the penalty function. This penalty accommodates the lasso ($\alpha = 1$), the ridge ($\alpha = 0$) and the elastic net [$\alpha \in (0,1)$], and permits the specification of different penalties for the coefficients $\{\phi_{ij}\}_{j\neq i}$ through the penalty scaling parameters $\{\omega_{ij}\}_{j\neq i}$.

Penalty scaling parameters are chosen according to the distances between spatial locations i and j, $j = 1, ..., n$. We consider several distance metrics in our analysis, including

- geographic distance, which measures the great-circle distance between two spatial locations and is denoted by $p = 1$,
- industrial distance, which measures the difference in the distribution of wages attributed to 19 industrial sectors between two spatial locations and is denoted by $p = 2$, and
- ethnic and racial distance, which measures the difference in the distribution of employees who are white, black, asian, and hispanic between two locations and is denoted by $p = 3$.

- population-weighted geographic distance, which measures the population-weighted great-circle distance between two spatial locations and is denoted by $p = 4$

If $\omega_{ik}^{(p)} < \omega_{ij}^{(p)}$, spatial location k is "nearer" to location i than location j; as a result, coefficient k is penalized less than coefficient j (less shrinkage is applied to coefficient k), increasing the likelihood that the estimate of coefficient k is nonzero and the estimate of coefficient j is zero (or arbitrarily close to zero).

1.7.5 Results

Figure 1.4 presents a forecast error comparison for the naive and penalized model for 380 metropolitan areas; each point represents a CBSA; for points that fall below the 45 degree line, the penalized model outperforms the naive model.

The inclusion of neighborhood information is an improvement over the naive model (the mean squared prediction error computed over an independent test sample is reduced by 18 per cent).

We conclude that extracting relevant spatial information through regularization can improve our understanding of labor market dynamics, generating better information for decision-makers.

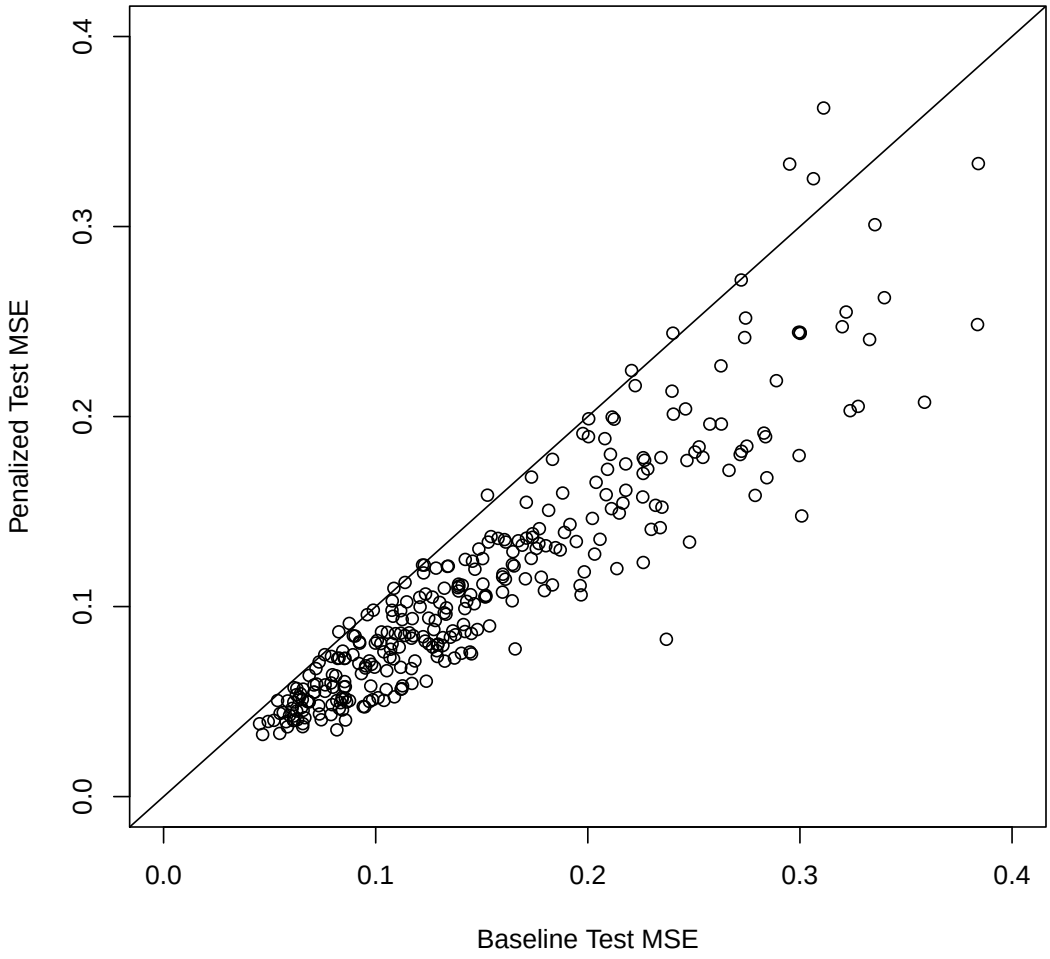


Figure 1.4: Forecast Comparison: Naive versus Penalized MSE Test Sample

BIBLIOGRAPHY

John M. Abowd and Lars Vilhuber. National estimates of gross employment and job flows from the quarter workforce indicators with demographic and industrial detail. *Journal of Econometrics*, 161(1):82–90, 2011.

John M. Abowd, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. The lehd infrastructure files and the creation of the quarterly workforce indicators. In Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts, editors, *Producer Dynamics: New Evidence from Micro Data*, chapter 5. The University of Chicago Press, 2009.

Jushan Bai and Serena Ng. Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163, 2008.

Timothy G. Conley and Giorgio Topa. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17(4):303–327, 2002.

Noel Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2011.

J Paul Elhorst. A spatiotemporal analysis of aggregate labor force behavior by sex and age across the european union. *Journal of Geographical Systems*, 10(2): 167–190, 2008.

Jianqing Fan, Jinchi Lv, and Lei Qi. Sparse high-dimensional models in economics. *Annual Review of Economics*, 3:291–317, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Clive Granger. Spatial data and time series analysis. *London Papers in Regional Science*, 1:1–24, 1969.

Enrique Lopez-Bazo, Tomas del Barrio, and Manuel Artis. The geographical distribution of unemployment in spain. *Regional Studies*, 39:305–318, 2005.

Kanti V Mardia, John T Kent, and John M Bibby. *Multivariate Analysis*. Academic Press, 1979.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL `http://www.jstatsoft.org/v39/i05/`.

James H Stock and Mark W Watson. *Handbook of Economic Forecasting*, chapter Forecasting With Many Predictors, pages 515–554. North Holland, 2006.

Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley, 2010.

William N. Venables and Brian D Ripley. *Modern Applied Statistics with S*. Springer, 2002.

CHAPTER 2

DETECTING AND REPAIRING DATA ERRORS IN THE PUBLIC USE QUARTERLY WORKFORCE INDICATOR DATA

2.1 Introduction

The Quarterly Workforce Indicators (QWI) (Abowd et al., 2009; Abowd and Vilhuber, 2011) provide quarterly statistics on employment, accessions, layoffs, hires, separations, full-quarter employment, job creations, job destructions, and earnings that can be queried by NAICS industry, geography (county, workforce investment board, and metropolitan and micropolitan area), and worker demographic characteristics (age, sex, education, and race/ethnicity).

Public use QWI distribution files (herein after referred to as the public use QWI data) are currently released with missing statistics and erroneous or contaminated statistics that may introduce bias when the data are used in empirical analyses of local labor market dynamics. *Missing statistics* refer to estimates that are missing for structural reasons (e.g., when the underlying data from which the statistic is computed are unavailable) or systemic reasons (e.g., when statistics are suppressed because they did not meet Census Bureau publication standards), and are flagged as missing or suppressed in the distribution files. *Erroneous or contaminated statistics* refer to statistics that have been contaminated by the late or inaccurate reporting of quarterly earnings.

Missing Data

QWI count data (accessions, separations, job creations, and job destructions) are released with corresponding status values, which indicate whether or not

a count is usable. Status values of -1 and -2 indicate that the count is missing for structural reasons, and a status value of 5 indicates that the count has been suppressed because it did not meet Census Bureau publication standards. Any count that has an associated status value of -1,-2, or 5 should be considered "missing" (or "unusable") and thus requires repair.

Data Errors

QWI data are derived from employer reports of quarterly earnings for the Unemployment Insurance (UI) program, which covers more than 96 percent of private sector employment. If firm-level reports of quarterly earnings are late, inaccurate, or in any way abnormal, the QWI are susceptible to discontinuities that require repair. For example, if a (sufficiently large) firm reports quarterly earnings late, we may see a spike in separations followed by a spike in accessions in the county or CBSA in which the firm operates; these spikes do not represent real labor market events (i.e., they do not represent a true birth/death when a firm enters/exits a market) and thus require repair. As another example, settlement of a class-action lawsuit might require a firm to make a one-time payment to current and/or previous employees, which could cause a spike in accessions followed by a spike in separations, if the firm is sufficiently large; again, these spikes do not represent labor market events and would require repair.

This paper develops a methodology for detecting and repairing missing and contaminated data, generating an improved (clean) version of the public use QWI data.

At present, the methods developed in this paper repair continuous-quarter

accessions to population rates (*capr*), continuous-quarter separations to population rates (*cspr*), job creation to population rates (*jcpr*), and jjob destruction to population rates (*jdpr*), for all industries, races, ethnicities, age groups, educations levels, and genders, summarized at the metropolitan and micropolitan geography level. The time frame considered spans 2000:Q3 to 2011:Q1.

The paper proceeds as follows: sections 2 and 3 describe our framework for analysis and statistical methods for outlier detection, section 4 outlines our proposed methodology for detecting and repairing QWI data errors, and section 5 presents our results.

2.2 Framework for Analysis

Our process for detecting and repairing missing and contaminated data involves two stages:

Stage 1: Detecting Missing Data and Data Errors

Detecting missing data is a trivial problem (all missing data are flagged as missing or suppressed); detecting data errors is not. Without information on events that cause data errors, such as information about late or inaccurate reports of quarterly earnings, we must approach the problem of detecting data errors by

1. detecting outliers[1] in the QWI time series, and

[1]We define an outlier as *an observation that appears to deviate markedly from other members of the sample in which it occurs* Barnett and Lewis (1994).

2. determining if the outliers are the result of errors, rather than true labor market events.

We can detect outliers in QWI time series using standard outlier detection methods. Determining if these discontinuities are the result of errors involves paired analysis of job flow data (e.g., paired analysis of accessions and separations data or paired analysis of job creation and destruction data) in addition to comparing the QWI discontinuities to discontinuities found in other, comparable data pertaining to local labor market dynamics.

Stage 2: Repairing Missing Data and Data Errors

We repair observations identified as missing or contaminated by forward (or backward) forecasting values based on previous (or future) values in the series, in addition to previous (or future) values of the corresponding nationwide series. For example, if r_t^s denotes the continuous-quarter accessions to population rate in period t in spatial location s, then a forward forecast is generated using the model

$$r_{t+1}^s = \phi_1 r_t^s + \phi_2 r_{t-3}^s + \phi_3 r_{t-4}^s + \phi_4 n_t^s + \phi_5 n_{t-3}^s + \phi_6 n_{t-4}^s + \eta_{t+1}^s, \quad t = 1, ..., \tilde{t} \tag{2.1}$$

where n_t he nationwide continuous-quarter accessions to population rate in period t, $\phi_j, j = 1, ..., 6$ are unknown parameters, η_{t+1}^s is an error term assumed to follow a normal distribution with mean zero and variance σ_η^2, and $\tilde{t}$ is the period that precedes the missing or erroneous observation that requires repair.

Similarly, a one-period-back forecast is generated using the model

$$r_{t-1}^s = \phi_1 r_t^s + \phi_2 r_{t+3}^s + \phi_3 r_{t+4}^s + \phi_4 n_t^s + \phi_5 n_{t+3}^s + \phi_6 n_{t+4}^s + \eta_{t+1}^s, \quad t = \tilde{t}, ..., T \tag{2.2}$$

where $\tilde{t}$ is the period that follows the missing or erroneous observation that

requires repair.

2.3 Outlier Detection Methods

We are faced with the task of detecting outliers that may represent data errors in the QWI data with no prior knowledge of which observations are, in fact, erroneous. This task is an unsupervised learning, or clustering, problem, in which we evaluate the data as a static distribution, pinpointing the most remote points and classifying them as potential outliers.

Methods for detecting outliers in an unsupervised context include, but are not limited to, statistical methods, distance-based methods (k nearest neighbors), parametric methods [minimum-value ellipsoid (MVE) estimation and convex peeling], semi-parametric methods (kernel methods and extreme value theory), and unsupervised neural methods.

Barnett and Lewis (1994) and Rousseeuw and Leroy (1996) provide texts on outlier detection in the statistics literature, while Hodge and Austin (2004) and Ben-Gal (2005) provide reviews of methods used in the artificial intelligence and data mining literature, and Tsay (1988) addresses the problem from a time series perspective.

In this paper, we evaluate statistical methods, turning first to a brief discussion of univariate versus multivariate methods.

2.3.1 Univariate versus Multivariate Methods

Outlier detection methods are generally suitable for either univariate or multivariate data. If an observation cannot be classified as an outlier when each variable is considered independently, then multivariate methods are appropriate.

In the case of QWI data, univariate analysis is sufficient for detecting "additive outliers" (outliers that affect only one observation of the series and not the future values).

Determining if these outliers are the result of errors, rather than the true birth or death of a firm, involves paired analysis of job flow data (e.g., paired analysis of accessions and separations data or paired analysis of job creation and destruction data) in addition to comparing the QWI discontinuities to discontinuities found in other, comparable data pertaining to local labor market dynamics.

2.3.2 Statistical Methods

The earliest univariate method for outlier detection assumes that the data are identically and independently distributed (i.i.d.) and that the underlying distribution is known (typically, it is assumed to be the Normal distribution $\mathcal{N}(\mu, \sigma^2)$). For any significance level $\alpha \in (0, 1)$, the outlier region of the $\mathcal{N}(\mu, \sigma^2)$ distribution is given by

$$\text{out}(\alpha, \mu, \sigma^2) = \{x_i : |(x_i - \mu)/\sigma| > z_{1-\alpha/2}\}$$

where z_q is the q quantile of the $\mathcal{N}(0, 1)$ distribution (Ben-Gal, 2005).

The outlier detection problem is one of identifying observations in the *outlier region* for a given sample $\{x_1, ..., x_n\}$. Typically, μ and σ are estimated by the sample mean and standard deviation,

$$\hat{\mu} = (1/n) \sum_{i=1}^{n} x_i, \qquad \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2},$$

and the significance level α is chosen by the researcher using receiver operating curves (ROC) to inform the selection process.

Since the sample mean and standard deviation can be greatly affected by outliers, they are often replaced with robust estimators, such as the sample median and the median of absolute deviations about the median:

$$\tilde{\mu} = \text{median}\{x_i\}, \qquad \tilde{\sigma} = \text{median}\{|x_i - \tilde{\mu}|\}.$$

2.4 Methodology for Detecting and Repairing QWI Data Errors

We propose the following methodology for detecting and repairing QWI data errors. This methodology assumes the analyst wishes to repair data for all CBSAs (metropolitan and micropolitan) for which data are available over a prespecified timeframe (in our case, 2000:Q3 to 2011:Q1).

2.4.1 Retrieving Standardized CBSA-Level QWI and QCEW Data

Retrieving Standardized CBSA-Level QWI Data

This section describes our process for retrieving standardized CBSA-level QWI data.

Retrieving CBSA-Level QWI Data

We retrieve QWI data from the "qwi_us_wia_metro_naicssec_pri" file, available for download at http://www.vrdc.cornell.edu/qwipu/. Records in this file are uniquely identified by state, year, quarter, ownercode/owner fm, agegrp/agegrp fm, sex/sex fm, metro/metrofm, and naicssec/naicssecfm. We consider records pertaining to privately owned firms, all sexes, all age groups, and all sectors (i.e., we consider records where ownercode="A05" and sex="0" and agegrp="A00" and naicssec="00") and summarize count data (continuous-quarter accessions, continuous-quarter separations, job creations, and job destructions) across states, generating a new file where records are uniquely identified by metro/metrofm, year, and quarter. Our procedure for retrieving CBSA-level QWI data involves these steps:

1. *Generate a list of CBSAs for which data are available over the desired time frame (2000:Q3 to 2011:Q1):* public use QWI data are derived from employer reports of quarterly earnings for the Unemployment Insurance (UI) program, provided by U.S. states, the District of Columbia, and territories that are in the Local Employment Dynamics (LED) partnership; the date at which U.S. states and territories joined the LED partnership varies; as a

result, for each CBSA included in our analysis, we consult the "dates.csv" file for the QWI release, for example

/data/clean/qwipu/state/data.R2012Q1/us/dates.csv,

to confirm that the state (or states) to which the CBSA belongs reported data from 2000:Q3 to 2011:Q1.

2. *Summarize CBSA data across state-level records:* Data for any CBSA that crosses state boundaries is captured in more than one state-level record; thus, to generate CBSA-level data, we summarize observations across state-level records.

3. *Compute CBSA-level continuous-quarter accessions and separations:* Continuous-quarter accessions are computed as $m - a - f$ and continuous-quarter separations are computed as $m - s - f$ where m denotes total employment, f denotes full-quarter (stable) employment, a denotes accessions, and s denotes separations.

4. *Flag missing data:* public use QWI data are released with data quality flags; the separations series (s) has data quality flag *status_s*, the accessions series (a) has data quality flag *status_a*, the full quarter employment series (f) has data quality flag *status_f*, the job creation series (jc) has data quality flag *status_jc*, and the job destruction series series (jd) has data quality flag *status_jd*. Continuous-quarter separations ($m - s - f$) are considered missing if either s or f has a status value of -2, -1, or 5. Continuous-quarter accessions ($m - a - f$) are considered missing if either a or f has a status value of -2, -1, or 5. Job creations (jc) are considered missing if jc has a status value of -2, -1, or 5. Job destructions (jd) are considered missing if jd has a status value of -2, -1, or 5. If a CBSA crosses state boundaries

and one of the state-level observations for the CBSA is considered missing, then the resulting CBSA-level observation generated in steps 2 and 3 should be flagged as missing.

Standardizing CBSA-Level Indicators

To detect outliers in the public use QWI "count" series, we first standardize the data with population estimates, generating rate data. For example, continous-quarter accessions to population rates (*capr*) are constructed using the following formula:

$$capr_t^s = \frac{m_t^s - a_t^s - f_t^s}{popest_{t-1}^s} * 100$$

where m_t^s denotes total employment, a_t^s denote accessions (new hires and recalls), and f_t^s denotes full quarter employment, in spatial location s during quarter t, and $popest_{t-1}^s$ denotes the end-of-quarter population estimate in spatial location s in quarter $t-1$.

Quarterly population estimates (*popest*) are derived from the annual CBSA population estimates available at http://www.census.gov/popest.

We first construct series of annual, county-level population estimates spanning 2000 to 2011 using the Vintage 2011 County Population dataset (with estimates for 2010 and 2011) and the County Intercensal Estimates (with estimates from 2000 to 2010). Intercensal estimates reconcile postcensal estimates (from 2000 to 2010) "with the 2010 Census counts and provide a consistent time series of population estimates that reflect the most recent census results."[2]

[2]See the Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010 at http://www.census.gov/popest/methodology/2000-2010_Intercensal_Estimates_Methodology.pdf

These annual estimates represent the number of people living in a county as of a specified point in time (usually July 1st) in a given year. Accordingly, we assume annual estimates represent the county population at the end of the second quarter and use linear interpolation to derive population estimates for the end of the first, third, and fourth quarters within the series.

We then summarize the quarterly, county-level estimates at the CBSA-level using the December 2009 metropolitan and micropolitan statistical areas and components definition file available at:

http://www.census.gov/population/metro/data/def.html.

Generating Standardized National Data

We repair missing and abnormal CBSA-level rate data using a forecast model that includes the nationwide standardized rate as a predictor.

Nationwide rate data is generated following these steps:

1. Generate a list of all CBSAs for which QWI data are complete and non-missing (i.e., with status values indicating that the counts are usable) over periods $t = 1, ..., T$.
2. Summarize QWI data for the list of CBSAs created in step 1, generating an aggregate, nationwide series identified by year and quarter.
3. Summarize population estimates for the list of CBSAs created in step 1, generating an aggregate, nationwide series identified by year and quarter.
4. Join the nationwide QWI and population estimates series, creating rate data of the form $qwi_t/popest_{t-1} * 100$ where qwi_t is an aggregate QWI observation in period t and $popest_{t-1}$ is the aggregate population estimate in

period $t-1$.

2.4.2 Repairing Missing QWI Data

For each metro-level standardized QWI rate that is considered missing, we forward or backward forecast its value based on previous or future values in the CBSA and corresponding nationwide rate series using the models specified in equations (1) and (2).

More specifically, for each CBSA-level rate series that has at least one missing observation, we:

1. Use trimmed least squares to estimate robust parameters and a set of robust residuals for models (1) and (2). Trimmed least squares is a robust estimation technique that involves (a) estimating models $j, j \in \{1,2\}$, using OLS and obtaining the residual, (b) ordering the residual and elimination observations associate with the three largest residuals, generating a trimmed data set, and (c) estimating model $j, j \in \{1,2\}$, using the trimmed data set, generating robust parameter estimates and a set of robust residuals.
2. Compute the distance between the time of the first missing observation in the series and the time of the first observation in the series and denote this by $dist1$.
3. Compute the distance between the time of the last missing observation in the series and the time of the last observation in the series and denote this by $dist2$.

4. If $dist1 > 5$, use model (1) to "forward forecast" all missing points for the CBSA, starting with the missing point that has the minimum time period. Repairs are made with the predicted value (generated with the robust parameter estimates from step 1) and a random draw from the normal distribution (with the mean and standard deviation computed from the sample of robust residuals generated in step 1).

5. Otherwise, if $dist2 > 5$, use model (2) to "backward forecast" all missing points for the CBSA, starting with the missing point that has the maximum time period. Repairs are made with the predicted value (generated with the robust parameter estimates from step 1) and a random draw from the normal distribution (with the mean and standard deviation computed from the sample of robust residuals generated in step 1).

6. If $dist1 < 5$ and $dist2 < 5$ (i.e., if there are missing observations at the beginning and end of the series), missing observations are repaired individually. That is, for the first missing observations, if the first, fourth, and fifth lagged values of this observation are non missing, we repair the observation using the forward forecast model. If the first, fourth, or fifth lagged value is missing, we check if the first, fourth, and fifth lead values are non missing and, assuming these lead values are non missing, we repair the observation using the backward forecast model. If it is not possible to repair an observation using either the forward or backward forecast model, we move to the next missing observation in the series, attempt to repair it, and once we have attempted to repair *all* missing observations in the series, we repeat the process, starting with the first missing observation in the (now partially repaired) series. The process is repeated until all missing observations have been repaired.

2.4.3 Detecting Outliers in the QWI Data

To detect outliers in our QWI rate series, we use the forward (backward) forecast models to compute residuals for the series and then use a univariate statistical method for detecting outliers in the residual. If the residual is deemed an outlier, the corresponding rate is also deemed an outlier.

For a given series $\{r_t^s\}_{t=1}^T$ and a significance level $\alpha \in (0,1)$, our process is as follows:

1. Find outliers in periods 5 through T using the forward forecast model:

   (a) Estimate model (1) using least trimmed squares (a process that is described in the previous section), generating robust parameter estimates for the model.

   (b) Compute the residual is model (1) for the full series using the robust parameter estimates.

   (c) Evaluate each residual $\{\eta_{t+1}^s\}$ and determine if it falls within the outlier region of the $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ distribution given by

   $$\text{out}(\alpha, \hat{\mu}, \hat{\sigma}) = \{\eta_{t+1}^s : |(\eta_{t+1}^s - \hat{\mu})/\hat{\sigma}| > z_{1-\alpha/2}\}$$

   where z_q is the q quantile of the $\mathcal{N}(0,1)$ distribution, $\hat{\mu}$ is the sample median,

   $$\hat{\mu} = \text{median}\{\eta_{t+1}^s\},$$

   and $\hat{\sigma}$ is the mean of absolute deviations about the median,

   $$\hat{\sigma} = \frac{1}{T-4} \sum_{t=5}^{T} |\eta_{t+1}^s - \hat{\mu}|.$$

(d) Flag the rate observations r^s_{t+1} that correspond with the outlying residuals (identified in step c) as outliers.

2. Find outliers in periods 1 through $T-5$ using the process described in step 1, except this time use the backward forecast model.
3. The unique set of points obtained in steps 1 and 2 are deemed to be *outliers*.

2.4.4 Determining if the Detected Outliers Represent True QWI Data Errors

Determining if detected outliers in the QWI data are the result of errors involves paired analysis of job flow data (e.g., paired analysis of continuous-quarter accessions to population and continuous-quarter separations to population rate data or paired analysis of job creation to population and job destruction to population rate data).

We apply the following rules to ascertain if outliers are the result of errors (an, therefore, need to be repaired):

- For each outlying observation in the *capr* series for spatial location s: if the *cspr* series in location s has an outlying observation within two periods of the time of the *capr* outlier, then the *capr* outlier is deemed a true outlier that requires reparation.
- For each outlying observation in the *cspr* series for spatial location s: if the *capr* series in location s has an outlying observation within two periods of the time of the *cspr* outlier, then the *cspr* outlier is deemed a true outlier that requires reparation.

- For each outlying observation in the *jcpr* series for spatial location *s*: if the *jdpr* series in location *s* has an outlying observation within two periods of the time of the *jcpr* outlier, then the *jcpr* outlier is deemed a true outlier that requires reparation.
- For each outlying observation in the *jdpr* series for spatial location *s*: if the *jcpr* series in location *s* has an outlying observation within two periods of the time of the *jdpr* outlier, then the *jdpr* outlier is deemed a true outlier that requires reparation.

2.4.5 Repairing QWI Data Errors

All data errors identified using the methodology describes in the previous sections are repaired using the same process as that described in the "Repairing Missing QWI Data" section of this paper.

2.5 Results

The methodology for detecting and repairing QWI data errors described in this paper results in a data set that excludes erroneous data resulting from late or inaccurate reports of quarterly earnings. Figures 2.1 through 2.3 portray QWI rate observations identified as outliers when $\alpha = .001$ and suggested repairs for those observations (red circles denote identified outliers, black points denote repairs for the identified outliers, and green points denote first-quarter observations).

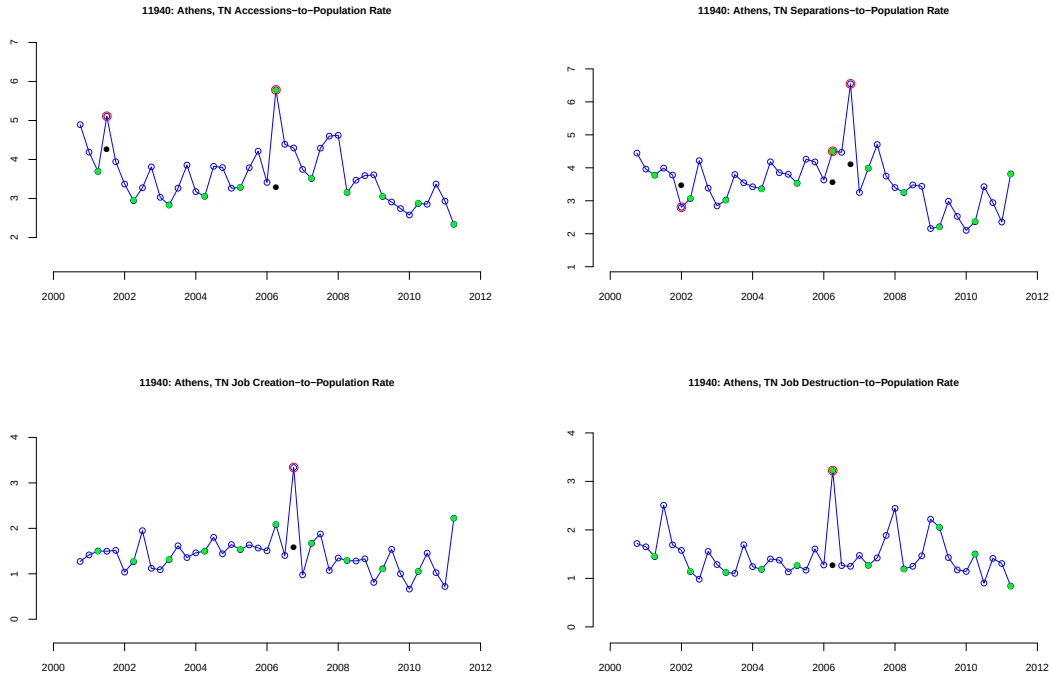


Figure 2.1: Outliers and repairs: Athens, TN

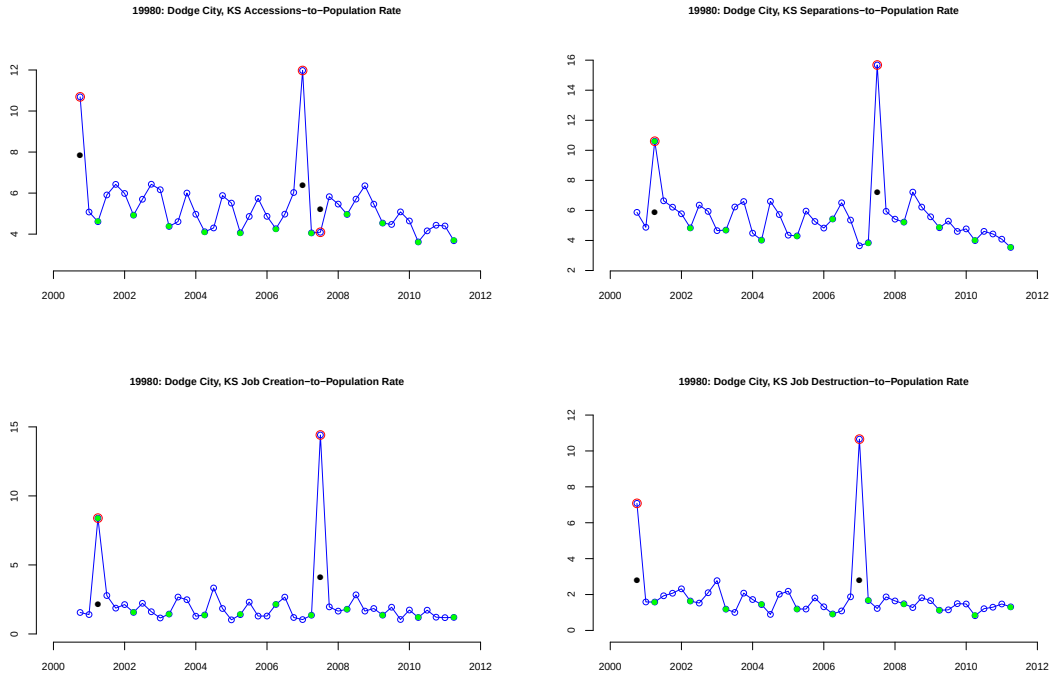


Figure 2.2: Outliers and repairs: Dodge City, KS

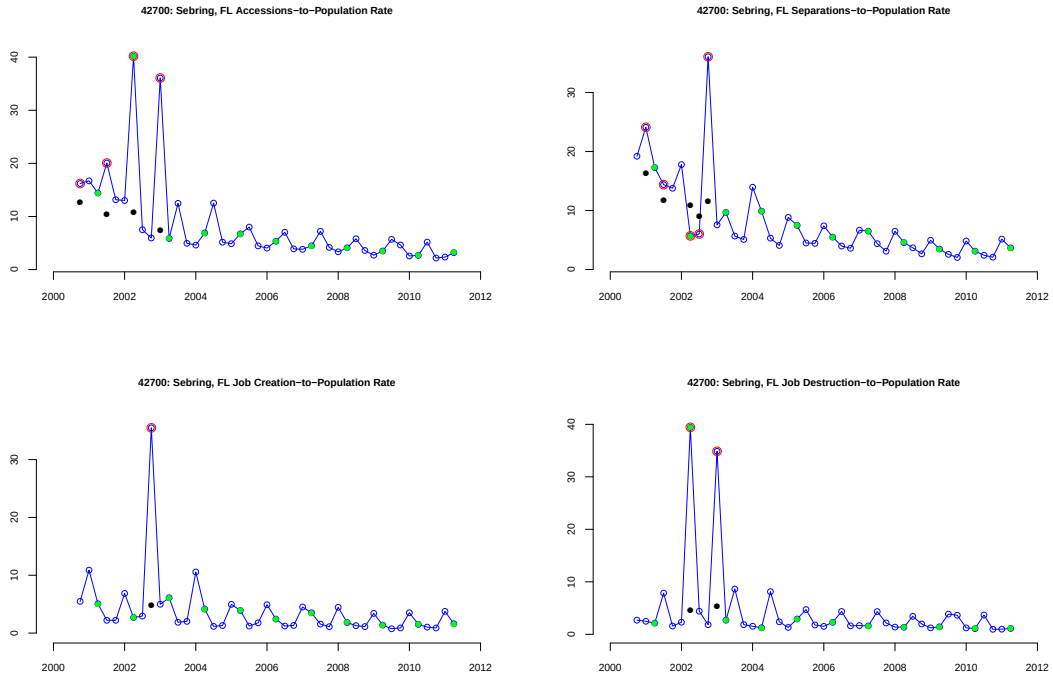


Figure 2.3: Outliers and repairs: Sebring, FL

At present, we do not have a sample for spatial rate data for which true outliers have been identified with certainty; thus, we are unable to analyze the extent of type I and type II errors (false positives and false negatives) when $\alpha = .001$. Future research in this area should include the generation of a sample where points are classified as being a true outlier (or not), permitting the researcher to use receiver operating characteristic curves to inform the choice of the discrimination threshold α. Future research may also include analysis of continuous-quarter accessions, continuous-quarter separations, job creations, and job destructions, broken down by age, race and ethnicity, and gender.

2.6 Appendix: QCEW and QWI Net Job Flow Comparison

Another method for confirming whether or not a detected outlier in the QWI data is a true outlier involves comparing the QWI data with comparable, external data, such as the Bureau of Labor Statistics's Quarter Census of Employment and Wages (QCEW), and determining if the difference between the data is exceedingly large (under the premise that a large difference suggests that the observation in at least one of the two sources is erroneous).

In the sections that follow, we describe a method for

1. computing the QCEW net job flow rates($(\tilde{e}_t - \tilde{b}_t)/popest_{t-1}$ where $\tilde{e}$ is the end of quarter employment count and $\tilde{b}$ is the beginning of quarter employment count),
2. computing the QWI net job flow rate ($(a_t - s_t)/popest_{t-1}$), and
3. detecting outliers in the difference between the net job flow rates.

2.6.1 Retrieving QCEW Data and Computing the QCEW Net Job Flow Rate

The QCEW program produces a comprehensive tabulation of employment and wage information for workers covered by State unemployment insurance (UI) laws and Federal workers covered by the Unemployment Compensation for Federal Employees (UCFE) program. Publicly available files include data on the number of establishments, monthly employment, and quarterly wages, by NAICS industry, county, and ownership sector, for the entire United States.

Our process for retrieving CBSA-level employment series and computing the QCEW net job flow rate follows:

1. From the qcew SAS dataset "bls_US_county_naicssec" (which uniquely identifies records by aggregation level, ownership code, naics sector, state, county, year, and quarter), retrieve county-level month1 employment counts, summarized across naics sectors, where aggregation level = "74" and the ownership code="5" (i.e., create county-level month 1 employment counts uniquely identified by state, county, year, and quarter).
2. Summarize the month 1 employment counts at the CBSA level and denote these month 1 employment counts as $\tilde{b}_t^s$ (the QCEW equivalent of the QWI beginning-of-quarter employment count b_t^s).
3. Create a lead month 1 employment count variable and denote it by $\tilde{e}_t^s$ (the QCEW equivalent of the QWI end-of-quarter employment count e_t^s).
4. Compute the change in employment in a given quarter as $\Delta\widetilde{emp}_t^s = \tilde{e}_t^s - \tilde{b}_t^s$, and the QCEW net job flow rate

$$qcew\_njfr \equiv \frac{\Delta\widetilde{emp}_t^s}{popest_{t-1}} * 100 = \frac{\tilde{e}_t^s - \tilde{b}_t^s}{popest_{t-1}} * 100$$

2.6.2 Computing the QWI Net Job Flow Rate

The QWI net job flow rate is computed as

$$qwi\_njfr \equiv \frac{a_t^s - s_t^s}{popest_{t-1}} * 100.$$

2.6.3 Detecting Outliers in the Difference Between the QCEW and QWI Net Job Flow Rates

We assume that, when the QCEW and QWI data are not erroneous, the following identity holds:

$$\frac{\Delta\widetilde{emp}_t^s}{popest_{t-1}} * 100 = \frac{\tilde{e}_t^s - \tilde{b}_t^s}{popest_{t-1}} * 100 \approx \frac{a_t^s - s_t^s}{popest_{t-1}} * 100.$$

As a result, the difference between the QCEW and QWI net job flow rates should be close to zero. Any difference that deviates markedly from zero is considered an outlier.

We use the statistical outlier detection method described previously to detect outliers in the difference between net job flow rates. We then compare whether the time periods of QWI rate outliers coincide with the time periods of the net job flow rate differences outliers. If the time periods of the outliers are concurrent, then we conclude that the QWI rate outliers are a true outliers (caused by true labor market events).

2.6.4 Results

Figures 2.4 and 2.5 portray results of the QCEW and QWI net job flow rate comparison. The top-left graph plots the accessions to population rate with identified outliers circled in red, the bottom-left graph plot the separations to population rate with identified outliers circles in red, the top-right graph plots QCEW and QWI net job flow rates, and the bottom-right graph plots the difference in the two net job flow rates with identified outliers circled in red.

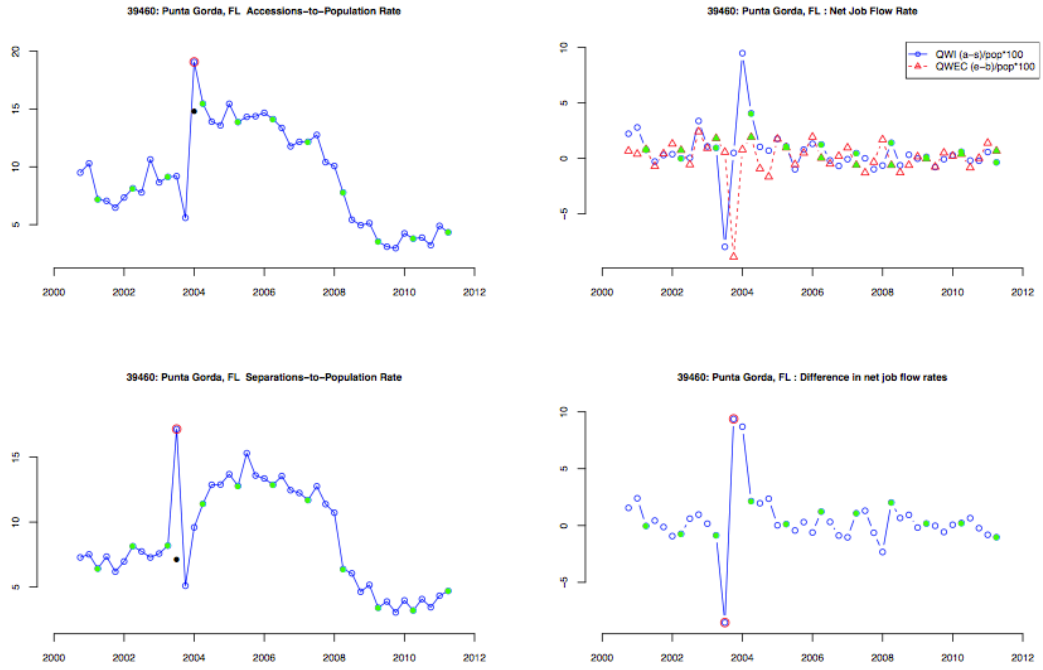


Figure 2.4: Net Job Flow Rate Comparison: Punta Gorda, FL: The separations rate outlier coincides with the net job flow rate difference outliers; the accessions rate outlier does not

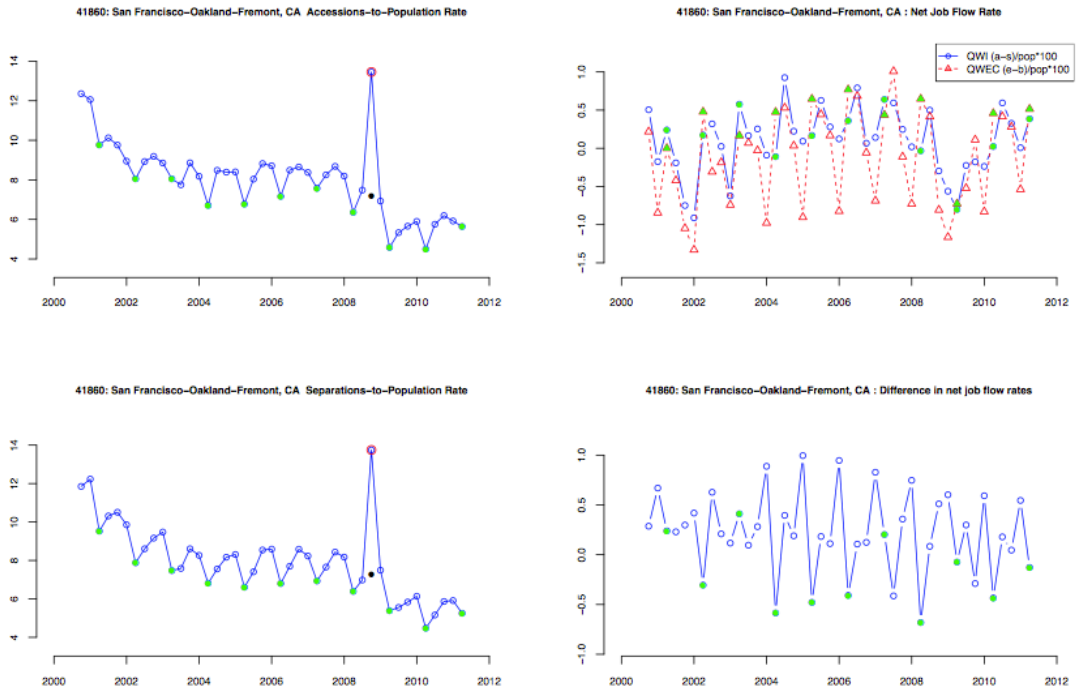


Figure 2.5: Net Job Flow Rate Comparison: San Francisco, CA: Neither the accessions nor the separations rate outliers coincide with the net job flow rate outliers

In many cases, the QWI outliers do *not* coincide with the outliers in the difference between the QWI and QCEW net job flow rates, as shown in figures 4 and 5. The mean and standard deviation of the QWI net job flow rate (taken across all time periods and all series) is .14 and 1.6, respectively, and the mean and standard deviation of the QCEW net job flow rate (taken across all time periods and all series) is -.005 and 2, respectively. This suggests that the two net job flow rates are not approximately equal and their difference is not centered on zero, which explains (at least in part) the results we see in figures 4 and 5.

Based on these preliminary results, the QCEW and QWI net job flow rate comparison is not recommended as a mechanism for determining if detected outliers in the QWI data are true outliers.

BIBLIOGRAPHY

John M. Abowd and Lars Vilhuber. National estimates of gross employment and job flows from the quarter workforce indicators with demographic and industrial detail. *Journal of Econometrics*, 161(1):82–90, 2011.

John M. Abowd, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. The lehd infrastructure files and the creation of the quarterly workforce indicators. In Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts, editors, *Producer Dynamics: New Evidence from Micro Data*, chapter 5. The University of Chicago Press, 2009.

Vic Barnett and Toby Lewis. *Outliers in Statistical Data*. Wiley, 1994.

Irad Ben-Gal. Outlier detection. In Oded Z. Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, chapter 7. Kluwer Academic Publishers, 2005.

Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1996.

Ruey S. Tsay. Outliers, level shifts, and variance changes in time seres. *Journal of Forecasting*, 7:1–20, 1988.

CHAPTER 3

REAL OPTIONS IN RESOURCE ECONOMICS

3.1 Introduction and Overview

Real options analysis is the application of option value techniques from finance to capital budgeting decisions affecting real investments (Trigeorgis, 1996). Capital investments and even short-term economic decisions often involve some degree of irreversibility and require financial commitment in the face of uncertainty. The construction of a hydroelectric project or the development of an offshore oil field requires specialized capital investments whose future return is uncertain because of volatile markets for electricity or oil. The decision to send a vessel out on a three-day fishing trip requires a commitment of resources with no guarantee of a profitable harvest. Timber that is cut and sold today will not be available if the price is higher tomorrow. An acre foot of water released from a reservoir for irrigation in spring will not be available in summer.

Decisions on the ultimate scale or size of a capital investment are often sequential. The real options approach allows one to determine the value of flexibility in both the scale and the timing of sequential investments. This flexibility is important when new information affecting the net value of a capital investment arrives continuously, periodically, or at random intervals in the future. The value of flexibility is not readily accounted for when discounted cash flow or other traditional capital budgeting techniques are used.

Weisbrod (1964) originally set forth the concept of option value for a park or hospital. Arrow and Fisher (1974), Henry (1974), Conrad (1980), and Han-

nemann (1989) explore option value as it might arise when making a decision to preserve or develop wilderness and how option value relates to the expected value of information.

Dixit and Pindyck (1994) identify irreversibility and uncertainty as key characteristics in determining the need for taking a real options approach to valuation. Those characteristics would seem to arise naturally in the field of resource economics, in which the investment to develop a natural resource, or even the timing of decisions to harvest or extract, involves varying degrees of irreversibility and uncertainty. The field of resource economics provided early examples in the development of real options analysis, including the opening and closing of a mine (Brennan and Schwartz, 1985), the value of an option to explore for oil on the continental shelf (Paddock et al., 1988; Cortazar and Schwartz, 1997), and the time to cut a stand of trees (Clark and Reed, 1989; Reed and Clark, 1990; Insley, 2002).

The seminal paper by Dixit (1989), although dealing with the entry and exit decisions of a generic firm, has important applications in natural resource economics, in which a decision to enter or exit may be irreversible. For example, the decision to build a large hydroelectric project or a nuclear reactor may be appropriately viewed as an irreversible entry decision. The decision to cut a stand of old-growth forest (Reed, 1993; Conrad, 1997) or to develop a wilderness area (Conrad, 2000) may be viewed as an irreversible exit decision from the state of old-growth forest or wilderness.

There have been many important applications of real options analysis in environmental economics. We view environmental economics as a separate subfield, but because of externalities [for example, greenhouse gases (GHGs)] as-

sociated with the extraction and use of nonrenewable resources (coal, oil, and natural gas), it is important to be aware of this literature. Pindyck (2007) provides a thorough review. Environmental policy with respect to climate change is a vexing problem because of the multiple sources of uncertainty and potential irreversibilities. Although this makes for rich applications of real options analysis, it means that the policy implications will depend importantly on the magnitude and timing of investments to reduce emissions and the future damage from climate change. Pindyck refers to the irreversibility of investments to reduce emissions versus irreversible climate change as opposing irreversibilities. Policy procrastination may be optimal if the cost of reducing emissions is high and the expected damage from climate change occurs in the distant future. Alternatively, irreversible investments to reduce GHGs would be made sooner, and at a larger scale, if the damage from climate change were large or likely to occur sooner or if the appropriate social rate of discount for climate change declined over time (Hepburn et al., 2009).

Kolstad (1996) and Fisher and Narain (2003) examine these opposing irreversibilities in two-period models. They find that investment irreversibility has a stronger effect than climate irreversibility and that the first-period reduction in emissions is modest. Using continuous-time models, Pindyck (1998, 2000, 2002) finds that if there is ongoing uncertainty over the benefits of reduced GHGs, along with uncertainty over the future evolution of GHG concentrations, then investments to reduce GHGs will be triggered at a higher GHG threshold and the scale of investment and level of emissions reduction at that threshold will be smaller.

The above papers do not consider the possibility of catastrophic climate

change. Clark and Reed (1994) consider a model in which consumption generates emissions that contribute to a stock pollutant, say the concentration of atmospheric carbon. Climate catastrophe is assumed to arrive according to a Poisson process, in which the Poisson parameter may be constant or increasing with the level of the stock pollutant. If the probability of catastrophic change increases with the concentration of atmospheric carbon, emissions are reduced now to lengthen the expected time to catastrophic change.

Tsur and Zemel (1996) construct a model in which climate catastrophe is triggered if and when the stock pollutant reaches or exceeds a critical but unknown threshold, or tipping point. This type of unknown, catastrophic threshold can lead to strongly precautionary behavior, in which the pollution stock is kept below a safe upper bound. Weitzman (2003), in his text *Income, Wealth, and the Maximum Principle*, develops a problem in which the arrival of information on the nature of global warming (whether it is good or bad) is a Poisson process and in which there is a given subjective probability that global warming will be found to be bad when conclusive information (the verdict) arrives. Given that you do not know when the verdict will arrive, or the nature of the verdict, how much forested land should you irreversibly develop today? (Forested land would be very beneficial if the news, when it arrives, is "global warming is bad.") Weitzman derives a critical threshold, reflecting real options value, which he calls the irreversible investment criterion (IIC). The linear structure of Weitzman's model results in an all-or-nothing rule so that all forest will be cut immediately if the IIC holds and no forest will be cut if the IIC does not hold. The point of the problem is to show that uncertainty in the arrival of information and the uncertain nature of that news (the verdict) create a strong incentive to conserve forested land today. However, if the expected arrival time for this

conclusive information is in the distant future, you will be more likely to cut forested land today.

Pindyck (2007) offers the following assessment.

> So, where does this leave us in terms of the policy implications for global warming? Should we adopt a stringent emissions reduction policy now, despite its cost, or go slowly and wait to learn more about the rate of global warming and its likely economic impact? To my knowledge, research to date does not give us the answer. Those studies cited above that ignore possible catastrophic impacts provide some evidence that we should move slowly. Those studies that do consider the possibility of catastrophic impacts suggest a more stringent emissions policy, but the catastrophic impacts in these studies are more or less assumed, rather than inferred from empirical evidence. Once again, we have a good understanding of economic theory, but a poor under- standing of its implementation in practice.

By limiting ourselves to the role of real options in resource economics (as opposed to environmental economics and global warming in particular) we avoid, for the most part, opposing irreversibilities and the attendant difficulties that arise in empirically determining their relative strength. This article is organized as follows. In the next section, we review the real options literature in forestry and give an example of a small woodlot owner contemplating timber harvest and land sale. In Section 3, we review the relatively sparse literature of real options in fisheries. We offer a simple model of Hamlet agonizing whether "to fish or not to fish." (Sorry, we couldn't resist.)

Section 4 reviews the literature on the use of real options to evaluate water resource development projects as well as options trading in markets for water. Water options are contingent claims for the existing supply of water. They are not real options because they are not being used to evaluate a physical, water resource investment, but if water markets that include water options exist, such markets can affect the value of water resource investments. Basically, a functioning water market with options trading may reduce the need for investments, like reservoirs, that would increase the total amount of water in a system. By better allocation of the existing supply of water (via options trading), there may be less of a need to expand water supplies, particularly during a drought.

Section 5 looks at nonrenewable resources and the option to extract. A one-shot extraction problem and a problem with continuous extraction and stochastically evolving reserves are solved. Section 6 examines the real options literature for conservation of natural environments and offers a problem identifying the net revenue threshold that would justify the loss of an amenity dividend and the option of future development. Section 7 concludes this article and suggests possible lines for future research.

Before we embark on the specific sections dealing with real options and different natural resources, it will be useful to establish some notation and forms for the commonly employed stochastic processes. We assume that net benefit, denoted $N = N(t)$, evolves stochastically according to three possible processes: geometric Brownian motion (GBM), given in Equation 3.1; a mean-reverting (MR) process, given in Equation 3.2; and a mean-reverting process with a Pois-

son jump (MRPJ), given in Equation 3.3.

$$dN = \mu N dt + \sigma N dz \tag{3.1}$$

$$dN = \eta(\bar{N} - N)dt + \sigma N dz \tag{3.2}$$

$$dN = \eta(\bar{N} - N)dt + \sigma N dz + \phi N ds \tag{3.3}$$

In Equation 3.1, μ is the mean drift rate in the proportional change in N. In Equations 3.1, 3.2, and 3.3, $\sigma > 0$ is the standard deviation rate, and $dz = \epsilon(t)\sqrt{dt}$ is the increment of a Wiener process, where $\epsilon(t) \sim N(0, 1)$. In Equations 3.2 and 3.3, $\bar{N}$ is the mean level to which net benefit tends to revert, and $\eta > 0$ is the speed of reversion. In Equation 3.3, the jump process, ds, is the increment of a Poisson process with a mean arrival rate given by $\lambda > 0$. It is assumed that dz and ds are independent. We use Equation 3.3 as a model for electricity prices, for which there is a possibility of a positive jump or spike in prices. We assume that

$$ds = \left\{ \begin{array}{c} 1 \text{ with probability } \lambda dt \\ 0 \text{ with probability } (1 - \lambda dt) \end{array} \right\} \tag{3.4}$$

Depending on our application, net benefit may be an ongoing process, or it may be a one-shot payout. For example, if an investment results in the construction of a hydroelectric project, $N = N(t)$ is the ongoing net benefit from a completed project. Alternatively, if we have a small, fixed reserve of oil that

can be instantaneously extracted and sold, then $N = N(t)$ may be a one-shot net revenue obtained at instant t. We think of $N = N(t)$ as net of any variable production or extraction costs. Depending on the application, we may also specify fixed costs to making an investment or to switching regimes.

3.2 Forestry

Many of the applications of real options analysis in forestry are concerned with the determination of the optimal age at which to harvest (or cut) a single, even-aged stand of trees. The decision to harvest is viewed as being irreversible. There is uncertainty over the future net revenue from harvesting, thus making the problem an ideal candidate for the real options approach to valuation.

A basic formulation may begin with the specification of an optimal stopping problem, in which, in any period or at any instant, the decision maker faces a binary choice of either harvesting the stand now (stopping) or waiting to harvest the stand at an optimal time in the future (continuing). Under certain regularity assumptions, there will be a single threshold that separates the stopping and continuation regions in such problems (see Dixit and Pindyck (1994)).

Clark and Reed (1989) and Reed and Clark (1990) derive harvest rules when the price of timber evolves according to GBM and the natural log of the volume of merchantable timber follows Brownian motion with drift. These researchers ignore harvesting and management costs. Thomson (1992) considers a model in which price follows a discrete-time analog of GBM and costs are specified as positive constants.Insley (2002) and Insley and Rollins (2005) examine the optimal price threshold to harvest an even-aged stand when price evolves ac-

cording to a GBM or a MR process while timber volume grows according to a deterministic, age-dependent function.

In the standard analysis of the optimal threshold to cut a single stand, most authors find that increases in the mean drift rate or standard deviation rate in the price process lead to increases in both the asset?s value prior to harvest and the critical threshold that would trigger harvesting. Thus, if prices are expected to trend upward, or if there is a greater possibility that the price will achieve an exceptionally high value, there is an incentive to delay harvesting—or more precisely, the expected time to reach the price trigger takes longer. Increases in the mean growth rate or variance rate lead to an increase in the value of the option to harvest. An increase in the discount rate decreases the value of the option to harvest.

Insley (2002) compares optimal harvest rules derived when the price of timber follows a MR process with those derived under a GBM process. When observed prices are below their long-run mean, the MR option value is considerably higher than the GBM option value, reflecting the fact that the price of timber will (eventually) rise in the MR scenario. The age of the stand and whether the volume of merchantable timber is growing rapidly (as with a relatively young stand of trees) or more slowly (as with an older stand of trees) also influence the critical threshold to cut. For white pine in Ontario, Canada, Insley (2002) finds that for stands 70 years or older, the trigger price under GBM is lower than under MR.

How does the risk of fire affect option value and the optimal age to cut? Reed (1984) and Insley and Lei (2007) analyze optimal resource management models that incorporate the risk of a catastrophic event specified as a Poisson process.

The probability that a catastrophic event will occur over the interval dt is given by λ, and if it occurs, the stock of the natural resource is reduced to zero. Reed showed that a Poisson fire risk increases the effective discount rate from δ to $(\delta + \lambda)$. Insley and Lei obtain this result in their model as well. In addition, with a risk of fire, the value of an option to delay harvest is reduced.

We now consider a simple model of a woodlot containing valuable timber but also providing an amenity flow (or dividend) when its owner walks the woodlot and "communes with nature." Suppose that the net revenue from the standing timber evolves according to Equation 3.1. If the woodlot owner ever cuts the timber, he will also sell the land. Suppose the amenity flow is constant, given by $A(t) = A > 0$, and that the value of the bare land, after harvest of the timber, is also a known constant given by $L > 0$. The amenity value is lost forever upon cutting of the timber and selling of the land. We assume that the woodlot owner?s rate of discount is denoted by δ and that $\delta > \mu \geq 0$. The last, and perhaps least realistic, assumption is that the woodlot owner has an infinite life span.

While optimally waiting to cut the timber and sell the land, the woodlot owner has a value function, $V(N)$, which must satisfy the Bellman equation,

$$\delta V(N) = A + (1/dt)E[dV] \tag{3.5}$$

where $E[\cdot]$ is the expectation operator. The left-hand side of Equation 3.5 can be interpreted as the interest income from selling the timber and land when the timber has a net revenue of $N = N(t)$. On the right-hand side of Equation 3.5, we have the benefit flow from continued ownership. This benefit flow comprises the amenity dividend, A, and the expected change in the value function over the infinitesimally small time step, dt.

With $N = N(t)$ evolving according to Equation 3.1, Ito's lemma implies that

$$(1/dt)E[dV] = \mu NV'(N) + (\sigma^2/2)N^2V''(N) \tag{3.6}$$

which upon substitution into Equation 3.5 yields

$$(\sigma^2/2)N^2V''(N) + \mu NV'(N) - \delta V(N) = -A. \tag{3.7}$$

Equation 3.7 is a second-order, nonhomogeneous, differential equation in the unknown value function, $V(N)$. The solution to Equation 3.7 is the function

$$V(N) = BN^\beta + A/\delta \tag{3.8}$$

where $B > 0$ is an unknown constant and

$$\beta = (1/2 - \mu/\sigma^2) + \sqrt{(1/2 - \mu/\sigma^2)^2 + 2\delta/\sigma^2} > 1 \tag{3.9}$$

The value function in Equation 3.8 has a logical economic interpretation. The term BN^β is the value of the (real) option to cut the timber and to sell the land. The term A/δ is the present value of the amenity flow if the woodlot owner never cuts and sells.

The one-time payout if the woodlot owner cuts and sells is $N + L$. There is a threshold, or trigger value, for N that would leave the woodlot owner indifferent between continued ownership and cutting the timber and selling the land. At this value for N,

$$BN^\beta + A/\delta = N + L \tag{3.10}$$

Equation 3.10 is termed the value-matching condition. The one-time payout of $N + L$ must compensate for the lost amenity value in perpetuity, A/δ, and the lost real option, BN^β.

There is a less intuitive condition termed the smooth-pasting condition, which requires continuity at the threshold where the woodlot owner is swapping the value function on the left-hand side of Equation 3.10 for the one-time

payout on the right-hand side of Equation 3.10. The smooth-pasting condition requires that the derivative of the left-hand side, with respect to N, equal the derivative of the right-hand side, with respect to N. In this problem, the smooth-pasting condition requires

$$\beta B N^{\beta-1} = 1 \tag{3.11}$$

Mathematically, Equations 3.10 and 3.11 constitute a two-equation system in two unknowns: the unknown constant, $B > 0$, and the unknown threshold N^*. We can solve Equations 3.10 and 3.11 and obtain an analytical solution for N^* that may be written as

$$N^* = \frac{\beta}{\beta - 1} \frac{A - \delta L}{\delta} \tag{3.12}$$

With numerical (parameter) values for A, L, μ, σ, and δ, we can calculate the numerical values for $\beta > 1$ and N^*. Then $B = 1/[\beta(N^*)^{\beta-1}]$.

Inspection of Equation 3.12 reveals that the woodlot owner would immediately cut the timber if $A/\delta \leq L$, in which case $N(t) > 0 \geq N^*$. This would occur if the perpetuity value of the amenity were less than the sales value of the land. As a numerical exercise, suppose that $A = \$1000$, $L = \$10,000$, $\mu = .01$, $\sigma = .1$, and $\delta = .06$. We can show that $\beta = 3$, $N^* = \$10,000$, and $B = 3.33333E - 09$.

3.3 Fisheries

Real options analysis has not featured prominently in fisheries economics or, more generally, in the literature dealing with the optimal management of a renewable resource. Nonetheless, certain investment decisions in fisheries management, such as the decision to exploit a fish stock under conditions of uncertainty or the decision to exit a limited-entry fishery, are obvious candidates for

real options analysis.

Li (1998) considers the option value of exploiting a cooperatively owned, license- restricted fishery where fishermen can deploy their fleet only once during a given season. He assumes that the fish stock follows a GBM process and that the dockside price of fish and harvesting costs are constants. He then formulates an optimal stopping problem. In this context, Li finds that stock size uncertainty leads to a more conservative fishing decision. Because the benefits of harvesting must equal or exceed both the costs of harvesting and the option to harvest at a later date, the real options approach has the fisher waiting until the stock is larger before exercising the option to fish.

Nostbakken (2006) analyzes the option to exploit a fishery when the decision maker is the sole owner of a small fishing ground. She assumes that the price and the stock of fish evolve according to known stochastic processes. Fishing is indicated by a binary variable, and there is a switching cost to go from inactive [$E(t) = 0$] to active [$E(t) = 1$] or from active to inactive. Similar to the discussion in Dixit (1989), the problem becomes an optimal entry-exit problem, but now with two stochastic processes affecting net revenue and option value. There will be two regime-switching curves in stock-price space, one for entering (activating a vessel or fleet) and the other for exiting (withdrawing a vessel or fleet). Nostbakken finds that pulse fishing is optimal, with the fishery often closed [$E(t) = 0$]. Sensitivity analysis shows that the maximum harvest rate has a large effect on the location of the switching curves, although the switching curves are less sensitive to the standard deviation rates.

Bosetti and Tomberlin (2004) evaluate an individual fisherman's decision to exit a limited-entry fishery in which the prices of fish and catch (the quantity of

fish landed) follow independent GBM processes. These researchers formulate an optimal stopping problem and use dynamic programming to determine the critical price and catch curve below which exiting is optimal and above which remaining active is optimal. The model correctly predicts 65% of 5059 observed exit decisions in a California salmon fishery.

The following model is consistent with the real options approach taken by Li (1998), Nostbakken (2006), and Bosetti and Tomberlin (2004). Consider a fisherman, call him Hamlet, who is pondering the question, "to fish or not to fish?" Let $X = X(t)$ be the stock of fish and $P = P(t)$ the dockside unit price for fish; both Ito variables evolve according to the equations

$$dX = [F(X) - qXE]dt + \sigma_X X dz_X \tag{3.13}$$

$$dP = \mu P dt + \sigma_P P dz_P \tag{3.14}$$

In Equation 3.13, $F(X)$ is a net growth function, and qXE is a fishery production function implying that harvest is the product of $q > 0$, a catchability coefficient, X the fish stock at time t, and fishing effort E, where $E = E(t) = \{0, 1\}$. If $E(t) = 0$, Hamlet has decided not to fish, whereas if $E(t) = 1$, Hamlet has gone fishing. The fish stock has an expected change per time step dt given by $F(X) - qXE$, but its evolution is stochastically influenced by the term $\sigma_X X dz_X$. Dockside price is assumed to evolve according to GBM (compare the form of Equation 3.14 with that of Equation 3.1). We also assume that the increments of the Wiener processes are independent so that $E[dz_X, dz_P] = 0$.

Fishing is costly. We assume that if $E(t) = 1$, the cost of the trip is $c > 0$.[2] With two Ito variables, and if we assume that Hamlet has an infinite life, the Bellman

equation requires

$$\delta V(X,P) = \max_{E(t)=\{0,1\}} [PqXE(t) - cE(t) + (1/dt)E\{dV\}] \tag{3.15}$$

where $\delta > 0$ is Hamlet?s discount rate. There will be a downward sloping curve in X - P space where Hamlet is indifferent between fishing and not fishing. After applying Ito's lemma to Equation 3.15, we define this curve by

$$H(X,P) = P - c/(qX) - V_X(X,P) = 0 \tag{3.16}$$

where V_X is the derivative of the unknown value function with respect to the fish stock. It is the shadow price of fish in the water.

In contrast to our woodlot owner, there is no analytic solution for the value function $V(X,P)$. Combinations of X and P that justify fishing lie above and to the right of $H(X,P) = 0$, as shown in Figure 3.1.

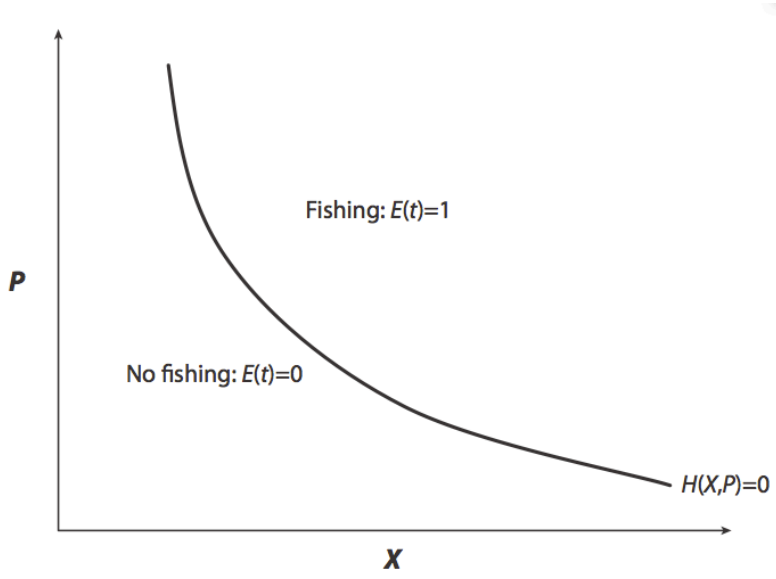


Figure 3.1: To fish or not to fish: the threshold curve $H(X,P) = 0$

3.4 Water Resources

In the literature on the economics of water resources, some authors have employed real options analysis to consider a farm's decision to invest in new, water-conserving irrigation technology. In addition, some authors have evaluated ways to improve water management and allocation via the use of options trading in water markets. We briefly review applications in both areas. Options trading in water markets is not an application of real options analysis because one is not evaluating a physical investment or an irreversible decision on using water today. Still, options trading may lead to a better allocation of the existing supply of water, reducing the need for investments in water resource development projects. Thus, we include this topic in this section.

Carey and Zilberman (2002) investigate a farm?s decision to invest in modern water-conserving irrigation technology when the farm has access to a stochastic supply of water available at no cost. This supply can be smoothed by buying or selling water in a spot market where the price of water follows a GBM. Carey and Zilberman observe a positive relationship between uncertainty in future water prices and the critical threshold required to induce investment, suggesting that policy makers should consider programs that reduce water price uncertainty if a higher rate of technology adoption is deemed socially desirable. Carey and Zilberman find that the availability of a spot market for water induces farms with an abundant (scarce) supply of water to adopt new technology at an earlier (later) date than they would if the market were not available. Thus, in the case of farms facing water scarcity, the introduction of a spot market for water will delay the adoption of a new, water-saving technology.

Seo et al. (2008) also investigate a farm's decision to adopt water-conserving irrigation technology. They focus on the declining Ogallala aquifer in the Texas High Plains, where groundwater pumping rates routinely exceed the recharge rate, with much of the water used for irrigation. They analyze the options to invest and disinvest in a modern irrigation system and derive critical thresholds in the form of agricultural output prices, which trigger both investment (entry) and disinvestment (exit). They find that, when the entry and exit decisions are considered simultaneously, the entry threshold is lower than it would be if they were to consider the decision to enter on its own. Furthermore, they find that the exit threshold is very low and not sensitive to parameter changes; thus, a farm using an old irrigation system is unlikely to divest of that system. As a result, to attain water-saving goals in the region, the authors suggest subsidizing the replacement of old irrigation systems with new water-conserving systems.

In a more general setting, Conrad and Kotani (2005) consider a farm's decision to invest in new technology intended to mitigate the effects of climate change. In areas of the United States that could formerly rely on natural precipitation, climate change may be inducing greater variability in precipitation, including a greater probability for protracted drought. They analyze the case in which crop biomass in a given season is stochastic, with a standard deviation rate that evolves stochastically as a result of climate change. Specifically, they assume that the seasonal standard deviation rate follows a discrete random walk with positive drift. Provided that the farm has the option to invest in new technology that will reduce the variance of the crop biomass, Conrad and Kotani pose the following question: How large must seasonal variation become before it is optimal to make the investment and adapt to climate change? In the infinite-horizon problem, suitable for a corporate or multigenerational family farm, a

single critical threshold induces the farm to adopt the new technology when the standard deviation rate in climatic variation exceeds that threshold. In the finite-horizon problem, suitable for a single farmer with no agricultural heirs, a schedule of critical thresholds triggers investment, with the critical threshold levels increasing as the farmer nears retirement. In both cases, the authors observe a negative relationship between the drift rate in the seasonal standard deviation rate of crop biomass and the critical threshold that triggers investment, indicating that an increased rate of growth in variability in harvest yields will induce farmers to adopt the new technology sooner.

Motivated by the increased demand for water in the growing urban areas in the southwestern United States, Michaelson and Young (1993) propose the introduction of water supply option contracts (WSOCs) as a way of meeting municipal water demand during droughts. In a case study, they examine an area along the Front Range in northeast Colorado. Assuming interruptible irrigation arrangements, they propose that a farmer (or group of farmers) has the ability to sell options for his or her water at a specified exercise price in the event of a drought. The holder of the option (an urban water user) would then have the right, but not the obligation, to buy water at the specified exercise price, should a drought occur. Michelsen and Young list conditions necessary for establishing WSOCs and then determine the value of such contracts. They conclude that the proposed contracts "have the potential to provide secure urban drought water supplies at a lower cost" than that resulting if urban users purchased the water rights directly or built new reservoirs.

Cui and Schreider (2009) analyze the water market in Victoria and New South Wales, Australia, where water is a traded commodity and supply and

demand determine the market price. Motivated by the issue of highly variable water prices, particularly during seasons with severe drought, they propose the introduction of water options, both call and put options, giving the holder the right, but not the obligation, to buy or sell water, thus providing greater assurance in both the price and the availability of water during a drought. Cui and Schreider discuss the benefits and challenges of introducing options to the water market and propose methods for modeling and computing option prices. Although the exact impact of introducing water options is unknown, they believe that water options are likely to provide farmers with greater income, reduced risk, and more choices in using their water allocation.

Without functioning water markets, a growing urban area in an arid climate would probably resort to building a reservoir to store the surface flow from a river system. The construction of a large reservoir can be viewed as an irreversible decision. If the future net benefits of the reservoir evolve stochastically, then real options analysis is appropriate. We here examine two water resource investments, the first for which net benefits from a completed project evolve according to Equation 3.1 and the second for which net benefits evolve according to Equation 3.3. As noted above, Equation 3.3 may be appropriate if the primary role of the project is the generation of electricity and if the price of electricity is subject to occasional price spikes.

Suppose that the water resource project requires a large fixed-cost investment denoted by $K > 0$. From Equation 3.1, net benefit (revenue less operating cost) evolves according to $dN = \mu N dt + \sigma N dz$. While one is optimally waiting to build the project, the option value term has the same functional form as in our woodlot problem and is given by BN^{β}. There is likely an amenity value to

the free-flowing river (prior to construction of the reservoir), but we ignore this possible dividend for now.

If and when the project is completed, it will provide a stream of stochastically evolving net benefits over an infinite horizon. The expected present value of this net benefit stream, if the project is completed at t when $N = N(t)$, is $N/(\delta - \mu)$. Recall that we are assuming $\delta > \mu$. This expected present value results from the fact that $N = N(t)$ is log-normally distributed with an expected value in $\tau \geq t$ given by $E[N9\tau)] = Ne^{\mu(\tau - t)}$. Then

$$\int_t^{\inf} Ne^{\mu(\tau-t)e^{-\delta(\tau-t)}} dt = N/(\delta - \mu) \tag{3.17}$$

To obtain the expected present value in Equation 3.17, one must incur the fixed cost $K > 0$ and kill the real option, BN^{β}. This leads to the value-matching condition

$$BN^{\beta} = N/(\delta - \mu) - K \tag{3.18}$$

and the smooth-pasting condition

$$\beta BN^{\beta-1} = 1/(\delta - \mu) \tag{3.19}$$

Equations 3.18 and 3.19 imply an analytic expression for the net-benefit threshold that would trigger construction, given by

$$N^* = \frac{\beta}{\beta - 1}(\delta - \mu)K, \tag{3.20}$$

where the expression for $\beta > 1$ is given in Equation 3.9.

Now suppose we are contemplating the same project but where the stochastic process describing the evolution of $N = N(t)$ is given by Equation 3.3, that is, a MRPJ. Let $V_0(N)$ denote the value function while one is optimally waiting. With

no amenity dividend to the free-flowing river, the Bellman equation requires

$$\delta V_0(N) = (1/dt)E[dV_0] = \eta(\bar{N} - N)V_0'(N) + (\sigma^2/2)N^2V_0''(N) + \lambda\phi N. \tag{3.21}$$

It is not possible to obtain an analytic expression for $V_0(N)$ that will satisfy Equation 3.21. With parameter values for δ, η, $\bar{N}$, σ, λ, and ϕ, numerical techniques will provide a functional approximation of $V_0(N)$.

Upon completion of the project, there is another value function, which we denote as $V_1(N)$, that must satisfy the Bellman equation,

$$\delta V_1(N) = N + (1/dt)E[dV_1] = (1 + \lambda\phi)N + \eta(\bar{N} - N)V_1'(N) + (\sigma^2/2)N^2V_1''(N). \tag{3.22}$$

Again, there is no analytic solution for $V_1(N)$. We know from our previous models that the value-matching and smooth-pasting conditions apply and can be used to find the critical threshold, N^*, where $V_0(N^*) = V_1(N^*) - K$ and $V_0'(N^*) = V_1'(N^*)$. In this problem, the MR process and the MRPJ process have opposite effects on N^*. If $\bar{N}$ is positive but modest in magnitude, whereas η is positive and of a magnitude that keeps $N = N(t)$ in a tight neighborhood around $\bar{N}$, then N^* may have to be high before the project is optimally undertaken. In contrast, the prospect of price spikes makes the project more attractive. If λ or ϕ are relatively large so that $\lambda\phi N$ has a large expected value, N^* would be lower and the project more attractive.

3.5 Nonrenewable Resources

In their seminal paper evaluating natural resource investments, Brennan and Schwartz (1985) initiate the discussion of when to open or close a copper mine

when the resource inventory is finite and the spot price evolves according to GBM. Cortazar and Casassus (1998) extend the Brennan and Schwartz model, assuming that price follows a MR stochastic process and evaluating the option to expand production (rather than the option to delay production). Slade (2001) develops a real options model for the decision to temporarily suspend and resume production in Canadian copper mines located in the provinces of British Columbia and Quebec. Slade assumes that both the spot price of copper and production costs follow MR stochastic processes whereas the quantity of remaining reserves follows either a stationary or a nonstationary stochastic process.

These authors begin their analysis by formulating a single- or multiple-state optimal-switching problem, which is solved by use of contingent claims analysis. In their analysis of a hypothetical mine, Brennan and Schwartz find that it is optimal to close the mine at low output prices and to open the mine at high output prices, assuming that the costs associated with closing and opening the mine are not exorbitant. Cortazar and Casassus's real options model concerns an actual copper mine in the United States with 12 years of reserves remaining; they find that it is optimal to accelerate (reduce) production in the early (later) years, due to the "lower net present value of later production in the presence of positive convenience yields." Furthermore, Cortazar and Casassus observe that lower quantities of initial reserves are associated with higher critical prices that trigger expansion. Slade compares a variety of alternative models in her analysis, distinguishing each model by the number of stochastic processes included in it and by assumptions regarding whether a particular process evolves according to a MR process or GBM. She finds that option values computed with nonstationary transition equations are systematically larger than those computed with

stationary equations, suggesting that GBM values may be misleading if the true price process is a MR process.

Kamrad and Ernst (2001) consider mining investments under conditions of market and process uncertainty. They assume that the output price evolves according to GBM and simulate operating risk by attaching a random multiplier to the output quantity, which reflects the "usable portion of output sold in competitive markets." They find that increases in the volatility of the price process lead to increalses in project value, whereas increases in the convenience yield lead to decreases in project value. They also observe that increases in production capacity are associated with increases in project value, but with diminishing marginal gains.

The above authors speculate that real options analysis is superior to traditional valuation methods, such as discounted cash flow and the net-present-value approach. However, they advise caution when specifying stochastic processes in a real options model, because stationary and nonstationary processes lead to significantly different results. Cortazar and Schwartz (1997) implement a real options model for valuing an undeveloped oil field when oil prices are stochastic and when production variables, such as remaining reserves and production costs, are deterministic. They assume that the oil spot price follows a mixed GBM and MR process and introduce a convenience yield that depends on the deviation of the spot price relative to a long-term average price. The solution identifies a critical spot price that triggers development of the field at which, once development is completed, production is initiated. They find that the option to delay investment accounts for a significant portion of the oil field's total value, particularly when oil prices are low. In addition, they observe that "the

critical price for development decreases with the available time to develop."

Paddock et al. (1988) use real options analysis to value leases for offshore petroleum when the net return from holding the lease follows a GBM diffusion process. Their model comprises a nested set of options that values the cash flow generated from three stages: exploration, development, and extraction. In an empirical example, these researchers note that exploration and development occur immediately if tracts are in the money, that is, the value of the developed reserve is known to exceed exploration and development costs. Tracts that are subject to higher investment costs, in contrast, exhibit positive option values for exploration and development.

We now consider two models, one in which the remaining reserves of a nonrenewable resource can be instantaneously extracted and sold (the one-shot model) and the other, a Hotelling-type model, in which stochastically evolving reserves are depleted over an infinite horizon. In the one-shot model, the choice variable is binary, whereas in the Hotelling-type model, the choice variable is a real-valued extraction rate.

In the one-shot model, let $R_0 = 1$ denote the normalized level of initial reserves. These reserves can be instantaneously extracted in their entirety at a cost of $K > 0$. The rate of extraction is a binary control variable, where $q(t) = 0, 1$. If and when $q(t) = 1$, $q(\tau) = 0$ for $\tau \neq t$. $P = P(t)$ denotes the unit price of $q(t) = 1$, and thus $P = P(t)$ is also gross revenue. We assume that $P = P(t)$ evolves stochastically according to GBM, where $dP = \mu P dt + \sigma P dz$. The owner of this one-shot reserve is assumed to have a discount rate of $\delta > \mu \geq 0$.

While one is optimally waiting to extract, the value of the option takes the

same form as in the woodlot owner?s problem but where $P = P(t)$ replaces $N = N(t)$ so that $V(P) = BP^{\beta}$. The exponent $\beta > 1$ is given by Equation 3.9. The net revenue obtained at instant t, when $q(t) = 1$, is $P - K$. The value-matching condition requires $BP^{\beta} = P - K$, whereas the smooth-pasting condition requires $\beta BP^{\beta-1} = 1=$. Together they imply the price trigger $P^* = \beta K/(\beta - 1)$.

Hotelling (1931) considered nonrenewable resource models in which the extraction rate was a continuous variable and it would be optimal to exhaust initial reserves over a finite or infinite time horizon (as opposed to an instant in the above one-shot model). If remaining reserves evolve stochastically, one needs to determine an optimal feedback extraction policy, $q^* = \phi(R)$, that will maximize the expected present value of remaining reserves. Pindyck (1980) considered problems of this sort. We present a simple model that has an analytic solution for the value function and optimal feedback extraction policy.

Now, instead of a single, one-shot net revenue given by $P - K$, we have a utility dividend given by $U = U(t) = q(t)^{1-\eta}/(1 - \eta)$, where $q = q(t)$ is the rate of extraction from remaining reserves, $R = R(t)$, and η is the Arrow Pratt measure of relative risk aversion. Remaining reserves are an Ito variable evolving according to

$$dR = -qdt + \sigma R dz \tag{3.23}$$

Equation 3.23 implies that future remaining reserves are not known with certainty but that the variance of expected future reserves declines as R declines. The optimization problem of interest seeks to

$$\begin{aligned} &\underset{\{q\}}{\text{maximize}} \;\; E\Big[\int_0^{\inf} [q^{1-\eta}/(1-\eta)]e^{-\delta t}dt\Big] \\ &\text{subject to } dR = -qdt + \sigma R dz, \;\; R(0) > 0 \;\; \text{given,} \end{aligned}$$

There is a value function , $V(R)$, that represents the expected present value of

remaining reserves, $R = R(t)$, when they are optimally extracted according to $q^* = \phi(R)$. The solution to this problem requires us to find the form for $V(R)$ and $q^* = \phi(R)$. The problem can be solved using dynamic programming and Ito's lemma. The Bellman equation requires

$$\delta V(R) = \max_q [q^{1-\eta}/(1-\eta) + (1/dt)EdV] \tag{3.24}$$

From Ito's lemma, we know that $(1/dt)EdV = -qV'(R) + (\sigma^2/2)R^2V''(R)$, and Equation 3.24 may be rewritten as

$$\delta V(R) = \max_q [q^{1-\eta}/(1-\eta) - qV'(R) + (\sigma^2/2)R^2V''(R)] \tag{3.25}$$

Maximizing the bracketed expression on the right-hand side of Equation 3.25 will give the optimal extraction rate as a function of $V'(R)$. Specifically, $q^* = [V'(R)]^{1/\eta}$. Substituting this expression back into the right-hand side of Equation 3.25, we can write the optimized Bellman as

$$\delta V(R) = \eta[V'(R)]^{-(1-\eta)/\eta}/(1-\eta) + (\sigma^2/2)R^2V''(R) \tag{3.26}$$

The critical question becomes, Is there an analytical expression for $V(R)$, with derivatives $V'(R)$ and $V''(R)$, that will satisfy Equation 3.26? The answer is yes. The value function $V(R) = \alpha R^{1-\eta}$ will satisfy Equation 3.26 when

$$\alpha = \left[\frac{(\sigma^2/2)\eta(1-\eta)^2 + (1-\eta)\delta}{\eta}\right]^{-\eta}(1-\eta)^{-(1-\eta)} \tag{3.27}$$

Then $q^* = [V''(R)]^{=1/\eta}$ will imply

$$q^* = [\delta/\eta + (\sigma^2/2)(1-\eta)]R \tag{3.28}$$

Equation 3.28 implies that optimal extraction is linear in remaining reserves, $R = R(t)$. For the feedback policy to make sense, $1 > [\delta/\eta + (\sigma^2/2)(1-\eta)] > 0$.

Equation 3.28 also implies that if either the discount rate, δ, or the standard deviation rate, σ, increases, the optimal rate of extraction, from a given level of $R = R(t)$, also increases. If η increases, the optimal rate of extraction decreases. Additionally, when $\sigma = 0$, $q^* = (\delta/\eta)R$, which is the optimal extraction policy in a deterministic world.

3.6 Conservation

As noted in the introduction, the seminal article by Arrow and Fisher (1974) was concerned with the amount of wilderness to develop when development was irreversible and the future value of wilderness was uncertain. In their model, quasi-option value resulted in a more conservative decision than if one had made the decision on the basis simply of expected benefits and costs. The literature subsequent to Arrow and Fisher (1974) continues to take the perspective of a government or social planner with the objective of allocating a natural resource or natural area between two competing and mutually exclusive uses: conservation and exploitation. This section describes the application of real options analysis to evaluate the optimal conservation of old-growth forest, wilderness areas, and species biodiversity.

Reed (1993) and Conrad (1997) address the decision of whether to conserve or harvest old-growth forest, recognizing that the habitat supporting old-growth trees often constitutes a unique or specialized ecosystem that supports other endangered species. Reed assumes that both future timber values (from harvest in the future) and amenity values (from a continued, standing forest) are stochastic. He finds that the optimal decision rule requires a comparison of the

"ratio of current timber value to the current, expected present value of amenity benefits" to a critical threshold ratio. Reed observes that traditional capital budgeting procedures, such as cost-benefit analysis, lead to premature harvesting; that is, they ignore the positive option of conserving the forest in the presence of stochastically evolving amenity values.

In Conrad's (1997) model, nontimber amenity values evolve stochastically according to GBM, and the net value of timber is deterministic. He derives the minimum critical amenity value needed to justify continued preservation. Increases in the net value of standing timber lead to increases in the minimum critical amenity value and make the cutting of old-growth timber more likely. Increases in the mean drift rate or standard deviation rate of the amenity flow lead to decreases in the minimum critical amenity value and make continued preservation more likely.

Chambers et al. (1994) consider forest preservation in a macroeconomic context, taking the perspective of a conservation organization that has acquired debt of a foreign country with the intention of swapping that debt for forest land preservation in what is known as a debt-for-nature swap. Recognizing that the value of the two assets (debt and nature) evolves stochastically, Chambers et al. evaluate the option to accept or reject the terms of the debt-for-nature swap in the time interval following the acquisition of the debt but before the swap. They view the debt-for-nature swap as an innovative market approach that has the potential to reduce Third World debt as well as tropical deforestation and its effects on the environment. They believe that a real options approach provides a useful framework for understanding the behavior of conservation organizations.

Conrad (2000) and Conrad and Kotani (2005b) consider the desirability and timing of decisions to extract resources from and/or to develop a wilderness area. In the latter paper, Conrad and Kotani evaluate the prospect of exploring for oil in the Arctic National Wildlife Refuge (ANWR) in Alaska, when the decision to develop incurs not only the cost of development, production, and transport but also the irreversible loss of amenity value. They evaluate the model when the price of oil follows a GBM or a MR process and derive the price that would trigger oil field development. In the MR scenario, the trigger price is considerably higher than in the GBM scenario.

Kassar and Lasserre (2004) evaluate species substitutability and biodiversity within a biological pool of species in an effort to understand whether substitutability reduces or increases the need for conservation. They formulate a model in which a decision maker continuously exploits the most usable species in the pool (for the provision of some product or service) and preserves the remaining unproductive species at a cost. The decision maker must choose between paying for the continued preservation of all unproductive (currently unused) species, which conserves the option of using them should they prove productive at a later date, or abandoning one or more of the currently unproductive species, which may result in their extinction. Kassar and Lasserre find that the option to substitute for currently productive species with a similar species is valuable, increasing the need for biodiversity conservation. Weitzman (1992) develops a theory of diversity in which an endangered species that is genetically close to an abundant, well-protected species is of lower diversity value than an endangered species with no genetically close relatives. The approach taken by Kassar and Lasserre is not at odds with that of Weitzman because commercial substitutability is entirely different from genetic distance.

Leroux et al. (2009) employ real options analysis to model the trade-off between land conversion and conservation. In their analysis, conservation benefits are a function of species richness and the economic value of biodiversity at the margin. In their model, an area of land that is currently used as a biodiversity reserve can be irreversibly converted to an alternative use, such as agricultural production. The option value of biodiversity results in a lower rate of conversion.

Conservation often provides an amenity dividend that is lost if an area is developed. Consider a reservoir that will result in the inundation of a large wilderness area. Let the amenity value of the wilderness be given by $A(t) = A$. For simplicity, we assume that A is a positive constant. Construction of the reservoir will cost $K > 0$. Upon completion, the reservoir will generate a stream of net benefits, $N = N(t)$, which evolve according to GBM, $dN = \mu N dt + \sigma N dz$d, with $\delta > \mu \geq 0$ and $\sigma > 0$.

The value function while one is waiting to build the dam has the same form as the value function of our woodlot owner waiting to cut his trees and sell his land. It takes the form $V(N) = BN^{\beta} + A/\delta$. See Equation 3.8, with $\beta > 1$ again defined by Equation 3.9. If and when the dam is completed, it will have a discounted expected value of $N/(\delta - \mu)$. The value- matching condition requires

$$BN^{\beta} + A/\delta = N/(\delta - \mu) - K, \tag{3.29}$$

whereas the smooth-pasting condition requires $\beta B N^{\beta-1} = 1/(\delta - \mu)$. Together, the value-matching condition and the smooth-pasting condition can be used to solve for $B > 0$ and the net revenue trigger

$$N^{*} = \frac{\beta(\delta - \mu)(A + \delta K)}{\delta(\beta - 1)} \tag{3.30}$$

By inspection of Equation 3.30, we can see that N^* will increase with increases in δ, A, or K and decrease with an increase in μ. Numerical analysis reveals that an increase in σ will decrease β and increase N^*

3.7 Future Research

The real options approach has been used to evaluate investment, harvest, and extraction policies for natural resources. The presence of irreversibility and uncertainty when these decisions are made makes real options analysis a compelling (and the correct) approach for valuation and management. Discounting expected net benefits to evaluate a risky, irreversible decision does not take into account the cost of "killing the option" and will result in the approval of projects that should not be undertaken or that have an excessive scale of development.

Applications to date have made extensive use of GBM and MR processes as a way to model the evolution of prices, output, net returns, or the amenity value of preserved wilderness. The value of a resource may be subject to occasional, discrete jumps, which implies that a Poisson process should be incorporated with another continuous process, resulting in a mixed stochastic process.

MR or MRPJ processes require numerical solution techniques to approximate value functions and optimal feedback policies and to estimate trigger values. Miranda and Fackler (2002) develop MATLAB utilities that may be used to computationally solve option value problems when analytic solutions do not exist. This is also the case for other discrete-time processes that may be more appropriate when studying price, net revenue, or the output of a specific natural resource. This reality suggests that resource economists will need to explore

the suitability of other stochastic processes and to develop additional (or customized) computational methods for determining the value and optimal use of these natural resources. In short, we need to expand our menu of stochastic processes and the computational techniques for solving valuation and management problems.

The application of real options analysis for conservation planning and for adapting to climate change is likely to be an important area of future research. Resource economists will also be needed to evaluate the external diseconomies associated with forestry and resource extraction and to determine the fair value of leases and rights to harvest natural resources on public lands. The public goods aspects of conservation, biodiversity, and climate change will be a high priority for economists, and real options analysis will be essential to the planning of investment and to the allocation of scarce resources over time.

BIBLIOGRAPHY

Kenneth J. Arrow and Anthony C. Fisher. Environmental preservation, uncertainty, and irreversibility. *Quarterly Journal of Economics*, 88:312–19, 1974.

Valentina Bosetti and David Tomberlin. Real options analysis of fishing fleets dynamics: a test. *Working Paper*, 51, 2004.

Michael J. Brennan and Eduardo S. Schwartz. Evaluating natural resource investment. *Journal of Business*, 58:135–37, 1985.

Janis M. Carey and David Zilberman. A model of investment under uncertainty: modern irrigation technology and emerging markets in water. *American Journal of Agricultural Economics*, 84:171–83, 2002.

Harry R. Clark and William J. Reed. The tree-cutting problem in a stochastic environment. *Journal of Economic Dynamics and Control*, 13:569–95, 1989.

Harry R. Clark and William J. Reed. Consumption/pollution tradeoffs in an environment vulnerable to catastrophic collapse. *Journal of Economic Dynamics and Control*, 18:991–1010, 1994.

Jon M. Conrad. Quasi-option value and the expected value of information. *Quarterly Journal of Economics*, 95:813–20, 1980.

Jon M. Conrad. On the option value of old-growth forest. *Ecological Economics*, 22:97–102, 1997.

Jon M. Conrad. Wilderness: options to preserve, extract, or develop. *Resource and Energy Economics*, 22:205–19, 2000.

Jon M Conrad and Koji Kotani. Adapting to climate change. *Working Paper*, 2005.

Gonzalo Cortazar and Jaime Casassus. Optimal timing of a mine expansion: Implementing a real options model. *Quarterly Review of Economics and Finance*, 38:755–69, 1998.

Gonzalo Cortazar and Eduardo S. Schwartz. Implementing a real option model for valuing an undeveloped oil field. *International Transactions in Operational Research*, 4:125–37, 1997.

J Cui and Sergei Schreider. Modelling of pricing and market impacts for water options. *Journal of Hydrology*, 371:31–41, 2009.

Avinash K. Dixit. Entry and exit decisions under uncertainty. *Journal of Political Economy*, 97:620–38, 1989.

Avinash K. Dixit and Robert S. Pindyck. *Investment Under Uncertainty*. Princeton University Press, 1994.

Anthony C. Fisher and Urvashi Narain. Global warming, endogenous risk, and irreversibility. *Environmental and Resource Economics*, 25:395–416, 2003.

M Hannemann. Information and the concept of option value. *Journal of Environmental Economics and Management*, 16:23–37, 1989.

Claude Henry. Option values in the economics of irreplaceable assets. *Review of Economic Studies*, 41:89–104, 1974.

Cameron J. Hepburn, Phoebe Koundouri, Ekaterini Panopoulou, and Theologos Pantelidis. Social discounting under uncertainty: a cross-country comparison. *Journal of Environmental Economics and Management*, 57:140–50, 2009.

Margaret Insley. A real options approach to the valuation of a forestry investment. *Journal of Environmental Economics and Management*, 44:471–92, 2002.

Margaret Insley and M Lei. Hedges and trees: incorporating fire risk into optimal decisions in forestry using a no-arbitrage approach. *Journal of Agricultural and Resource Economics*, 32:492–514, 2007.

Margaret Insley and Kimberly Rollins. On solving the multi rotational timber harvesting problem with stochastic prices. *American Journal of Agricultural Economics*, 87:735–55, 2005.

Bardia Kamrad and Ricardo Ernst. An economic model for evaluating mining and manufacturing ventures with output yield uncertainty. *Operations Research*, 48:690–99, 2001.

Charles D. Kolstad. Learning and stock effects in environmental regulation: the case of greenhouse gas emissions. *Journal of Environmental Economics and Management*, 31:1–18, 1996.

E Li. Option value of harvesting. *Marine Resource Economics*, 13:135–42, 1998.

AM Michaelson and RA Young. Optioning agricultural water rights for urban water supplied during drought. *American Journal of Agricultural Economics*, pages 1010–20, 1993.

Mario J. Miranda and Paul L. Fackler. *Applied Computational Economics and Finance*. MIT Press, 2002.

Linda Nostbakken. Regine switching in a fishery with stochastic stock and price. *Journal of Environmental Economics and Management*, 51:231–41, 2006.

James L. Paddock, Daniel R. Siegel, and James L. Smith. Option valuation of claims on real assets: the case of offshore petroleum leases. *Quarterly Journal of Economics*, 103:479–508, 1988.

Robert S. Pindyck. Uncertainty and exhaustible resource markets. *Journal of Political Economy*, 88:1203–25, 1980.

Robert S. Pindyck. Sunk costs and sunk benefits in environmental policy. *Working Paper*, 95-003, 1998.

Robert S. Pindyck. Irreversibilities and the timing of environmental policy. *Resource and Energy Economics*, 22:233–59, 2000.

Robert S. Pindyck. Optimal timing problems in environmental economics. *Journal of Economic Dynamics and Control*, 26:1677–97, 2002.

Robert S. Pindyck. Uncertainty in environmental economics. *Review of Environmental Economics and Policy*, 1:45–65, 2007.

William J. Reed. The effects of the risk of fire on the optimal rotation of a forest. *Journal of Environmental Economics and Management*, 11:180–90, 1984.

William J. Reed. The decision to conserve or harvest an old-growth forest. *Ecological Economics*, 8:45–69, 1993.

William J. Reed and Harry R. Clark. Harvest decisions and asset valuation for biological resources exhibiting size-dependent stochastic growth. *International Review of Economics*, 31:147–69, 1990.

S Seo, E Segarra, PD Mitchell, and DJ Leatham. Irrigation technology adoption and its implication for water conservation in the texas high plains. *Agricultural Economics*, 38:47–55, 2008.

Margaret E. Slade. Valuing managerial flexibility: an application of real-option theory to mining investments. *Journal of Environmental Economics and Management*, 41:193–233, 2001.

Thomas A. Thomson. Optimal forest rotation when stumpage prices follow a diffusion process. *Land Economics*, 68:329–42, 1992.

Lenos Trigeorgis. *Real Options: Managerial Flexibility and Strategy in Resource Allocation*. MIT Press, 1996.

Yacov Tsur and Amos Zemel. Accounting for global warming risks: resource management under event uncertainty. *Journal of Economic Dynamics and Control*, 20:1289–1305, 1996.

Martin L. Weitzman. *Income, Wealth, and the Maximum Principle*. Harvard University Press, 2003.