Correlating Community Specific Rural Diesel Fuel Prices with Published Indices of Crude Oil Prices, And Potential Price Projection Applications

June 2015

Prepared for:

Alaska Energy Authority

Prepared by:

Dominique Pride
Matthew Snodgrass
Antony Scott

Table of Contents

1. Introduction

Most rural Alaskan communities rely on diesel for both space heating and electricity generation. This can create economic hardship for rural communities when the price of diesel is high and create incentive for communities to look for alternatives. Economic evaluation of alternatives must necessarily be made, at least in part, against the value of fuel oil displaced. Unfortunately for community or state-based energy managers, community-specific forecasts of fuel oil are not generally publicly available.

In 2008, the Alaska State Legislature established the Rural Energy Fund to provide competitive funding for qualifying renewable energy projects throughout Alaska (Alaska Energy Authority, 2015). One criterion for proposed renewable energy projects is an economic analysis to assess the feasibility and fuel savings of proposed projects. This particular program relies on fuel price projections to conduct these assessments.

This report develops rural fuel price models to facilitate logically consistent projections of delivered prices across communities. The parsimonious approach relies primarily on published index crude oil prices to explain variation in delivered prices of fuel. Past work by the University of Alaska Anchorage's Institute of Social and Economic Research (Szymoniak, Fay, Villalobos-Melendez, Charon, & Smith, 2010) and our own interviews with fuel distributers and rural electric utility managers confirm that most fuel supply contracts between rural utilities and distributers tie the cost of delivered fuel to a window around the date that refined product is lifted from the refinery. Published index product prices are used in contract terms, presumably to allow buyers to confirm that delivered prices reflect the price terms negotiated with the distributer. An additional markup above the distributer's cost of fuel is negotiated to compensate for costs of delivery and a margin of profit. The structure of such contracts ensures that risks of future changes in refinery product costs are ultimately borne by the customer.

It can be readily confirmed that the price of diesel fuel at the refinery on any given day is tightly correlated with contemporaneous published crude oil prices. Distributer markup is a negotiated term where the determinant is not transparent, but is presumably affected by a host of community-specific factors including creditworthiness, distance from refinery, total fuel demand, relative bargaining power, and so on. It stands to reason that precision in prediction of community-specific distributer markup – let alone potential causal determinants that explain why community markup varies across communities – will be enhanced with precision in the estimated date at which refined product is lifted at the refinery. The central effort and contribution of this study is to seek to better isolate the 'correct' explanatory crude oil prices that affect refinery product prices, so that parameter estimates of determining community-specific markups can made more precise. The actual correct crude oil prices are the prices obtaining the relevant contractual window around refinery lift.

Rural utilities participating in the Power Cost Equalization (PCE) program submit fuel invoices to the Regulatory Commission of Alaska (RCA). Invoices were retrieved from the RCA's online

document library, and from paper records, and assembled into a database which was used to build rural fuel price models for barge, air, and road deliveries (Regulatory Commission of Alaska, 2015). The Least Squares Dummy Variable estimator method was used to estimate community-specific models categorized by mode of fuel transportation. Crude oil prices can then be plugged into these models to project each community's delivered fuel prices. The resulting estimates of delivered fuel prices can be used to assess the comparative economic feasibility of future renewable energy projects in rural communities.

The remainder of the report is organized as follows: Section 2 on methods and procedures describes the data, estimation of the date of fuel lift, the fixed effects least squares dummy variables estimator, the estimation of the fuel price models, and the price projections. Section 3 covers the results of the estimated fuel price models and the price projections. Section 4 is discussion, followed by the conclusion in Section 5. The appendices provide instructions for updating both the database and the estimated models.

2. Methods and Procedures

Data

Data for this project was collected from several different sources. Fuel invoices were collected from the RCA. Data for measuring the distance between a community and the refinery was calculated using the National Oceanic and Atmospheric Administration's Distance between United States Ports (2012) and the distance tool in Google Earth. Data for annual fuel consumption by a community's electric utility was collected from the Power Cost Equalization Utility Statistics[1]. Brent crude oil prices, expressed in nominal dollars per barrel, from 2000 to 2012 were purchased from Platts. Brent crude prices for 2013 were collected from Alaska Department of Revenue Tax Division (2015). All price data was adjusted for inflation using the United States Bureau of Labor Statistics (2015) Historical Consumer Price Index for All Urban Customers (CPI-U).

Fuel invoices are submitted to the RCA by utilities participating in the PCE program. The RCA maintains an online database where these fuel invoices can be accessed by the public. In-person visits were made to the RCA office to collect additional fuel invoices that were not posted to the online database.

Each invoice collected for a community is an observation in the (uncleaned) dataset. This project is only concerned with invoices for public utilities. For each invoice, the price per gallon of

[1] Monthly utility fuel use was summed for each fiscal year. Where three or fewer months of fuel data were missing, annual fuel consumption was imputed using average efficiency based on the fuel use and kilowatt hours produced in the non-missing months for a fiscal year. Where three or fewer months of fuel data was missing and kilowatt hours produced were also missing, the reported fuel usage for remaining months was summed over the fiscal year and used as annual consumption. If more than three months of fuel use data were missing, annual consumption was left blank.

delivered fuel, quantity of the delivery, distributor, purchaser, fuel type[2], payment terms, invoice date, delivery date, and mode of delivery were recorded.

Collecting data from actual invoices offers multiple benefits. Examining an invoice often allows one to determine the mode of delivery. For example, if an air freight company is listed on the invoice as the fuel seller, then the mode of transportation for that invoice is air. This is helpful because some communities receive both barge and air deliveries. Collecting data from individual invoices can aid in estimating the date of lift from the refinery because the delivery date is usually listed on the invoice. Also, looking at the actual invoice can lend insight to complex local arrangements. Some transactions are not arm's length but instead are between connected entities. For example, sometimes the local fuel merchant and the electric utility are not independent from one another.

The panel dataset covers fuel deliveries for 153 Alaskan communities from 2000 through 2013. The dataset is irregular due to the time spacing between observations varying from community to community. This is because communities do not all receive their fuel shipments on the same day. The dataset is unbalanced because not all communities have the same number deliveries, and some communities are missing observations for a given year. There are several possible explanations for missing observations. It is possible that a community did not submit fuel invoices to the RCA for a given year. It is possible that a community did submit invoices but the invoices were lost in the mail or lost once they arrived at the RCA and were never scanned into the database. For this project, it is assumed that all missing observations are missing at random, meaning that missing observations do not depend on the unobserved data (Allison, 2009).

In many cases, the submitted fuel invoice data had to be cleaned due to number of issues. For example, in some cases, the relevant mode of delivery is not the same as what is recorded on an invoice. On Prince of Wales Island, fuel invoices from the distributer sometimes appear to be "trucked" when they generally must be first be transported by barge from Seattle and then trucked to the various communities on the island. Invoices where the mode of delivery could not be deduced with some measure confidence were removed from the dataset. Similar issues arise for some barge communities that have many fuel invoices for small quantities of fuel even in months when rivers and ports are frozen. Any fuel invoice for an ice-bound barge community with a delivery date in December, January, or February was removed from the dataset because a barge cannot navigate ice-bound waters. It is assumed that these deliveries are pulled from a storage tank and trucked to the local power plant.

In cases where a community had a string of invoices over time with a constant price, these observations were collapsed into a single (the first) observation. It is assumed that the string of deliveries with a constant price result from a single fuel contract. Therefore, the first delivery

[2] The fuel type is not always listed on the invoice. Many observations are not associated with a fuel type. Rather than discard all observations without a fuel type, all fuel types are included in the model.

date in the string of deliveries with a constant price must be the closest to the contract date on which the fuel price was determined. Collapsing observations in this manner was necessary for road communities and some South East communities.

In general, the invoices lack the date of fuel lift from the refinery. If it is assumed that fuel is not stored before it is sold, these lift dates can be estimated. However, there is no way to know for certain that fuel was not stored before being sold because invoices do not indicate whether or not fuel was pulled from storage. This is problematic because fuel is priced on the date of lift from the refinery. There is no way of estimating the date of lift of fuel that has been pulled from storage since one cannot know when it was put into the storage tank. For this project, it is assumed that fuel was not delivered out of storage. As noted, some of the data cleaning procedures adopted were designed to improve the validity of this assumption.

In total, there are 6,823 observations in the cleaned dataset. The dataset is split into sub-sets based on mode of transportation. It is assumed that different modes of delivery follow different price processes. This is because of the variation in transportation costs and logistical difficulties across modes of transportation. The data were grouped into barge, air, and road deliveries. Barge deliveries were further divided into ice-bound communities and ice-free communities. While communities with ice-free harbors have access to barge-delivered fuel year-round, communities with ice-bound harbors can only receive barge deliveries in warm, ice-free months. There are 96 ice-bound barge communities, 33 ice-free barge communities, 32 air communities, and 14 road communities. There is no overlap between ice-bound barge, ice-free barge, and road communities. However, there is overlap between ice-bound barge and air communities. It is possible for a community to run out of fuel after the water freezes, or a river may be too shallow to navigate during the ice-free months in some years. In these situations, a community that would normally have fuel delivered by barge must instead receive air deliveries. Separate models were estimated by each mode of transport.

Estimating Lift Date

As mentioned previously, the delivered price of fuel is determined on the date of lift. Knowing the date of lift allows the price of crude oil on the date of the fuel contract to be identified. For air and road deliveries, it is assumed that all deliveries are made on the same day the fuel was lifted[3]. The delta between fuel lift and delivery for barge deliveries to a given community is both variable across communities and stochastic within a community; it is affected by the distance between the community and refinery, weather conditions throughout the barge's voyage, the speed at which the barge is traveling, and the number of delivery stops the barge makes before reaching a given community, among other factors. Unfortunately, lift dates are not listed on the majority of invoices. However, Alaska Village Electric Cooperative (AVEC) provided lift dates

[3] This assumption is reasonable given actual transit times. A plane can fly across Alaska in a single day. Additionally, given the small distribution area available for road deliveries, trucked deliveries can arrive the same day they are lifted.

for all deliveries to their communities between 2009 and 2013. Additionally, during 2007 and 2008, Crowley invoices recorded lift dates for a portion of AVEC deliveries. This dataset was used to estimate a model to predict the number of days between lift and delivery for all ice-bound barge communities. In doing so it has been assumed that variation in lift date is a function of inherent logistics, rather than the identity of the customer.

The delta between lift and delivery dates for fuel deliveries to each community was estimated as

$$\ln(\delta) = \beta_0 + \beta_1 Distance + \beta_2 YukonRiver$$

where $\ln(\delta)$ is the natural log of the number of days between lift and delivery, *Distance* is the number of miles between the Kenai refinery and a community via the ocean including the distance from the mouth of the river to a community, except for Yukon River communities in which case, *Distance* is the number of miles between the community of Nenana and a community along the Yukon River. *YukonRiver* is a dummy variable indicating a whether or not a community is located along the Yukon River. Communities along the Yukon River receive fuel shipments from barges that depart from Nenana[4]. Efforts to model the delta as a function of other variables – e.g. specific distance of ocean versus river transport, total quantity consumed, measures of the community's credit-worthiness – were both exhaustive and proved fruitless.

Each community's "expected delta" between lift and delivery date allows a mapping from the known delivery date to Brent price. However, what is really desired is the expected Brent price that corresponds to the delivered price, as this is the best measure of the (uncertain) crude oil price used by the refinery to make product. The expected Brent price is not the expected delta, mapped to singular Brent price. Instead, the expected Brent price is a probability-weighted Brent price, where the weights reflect uncertainty in the true delta.

The calculation of the expected Brent price that corresponds to the product date of lift is straight-forward. The estimated delta between lift date and delivery date defines the lower and upper bounds of the 95% confidence interval around the expected lift-delivery delta. Each date within the confidence interval is mapped to its corresponding Brent crude oil price, and each of these prices is weighted by the probability of such a delta. The sum is just the expected Brent price at the time of lift corresponding to the known date of delivery. For example, if the delta between lift and delivery was estimated to be five days and had a confidence interval with a lower bound of three days and an upper bound of seven days, the Brent prices corresponding to three days to seven days before the delivery date would be weighted by their respective probabilities from the normal probability distribution and then summed resulting in a single weighted Brent price for that observation.

[4] Communities at the mouth of the Yukon River receive fuel shipments via ocean-going barge. These communities were grouped with the coastal communities instead of the Yukon River communities.

Barge communities that remain ice-free during the winter should be expected to have different deltas between lift and delivery than ice-bound barge communities given significant differences in logistics. However, none of the AVEC communities are ice-free, making predictions based on AVEC data problematic. Accordingly, a separate statistical model was developed that exploits the time dependency of price to develop an estimate of the probability distribution of the uncertain time between lift and delivery for each ice-free community.

Transit time between lift and delivery is subject to a number of factors likely to induce variation across deliveries. For example, poor weather is likely to increase the delta between lift and delivery above what is typical for the community. Alternatively, there may be countervailing forces for a given delivery that reduce the delta between lift and delivery below what is typical. Smoothing across a number of days was used in the model to account for this unobserved variability. Put differently, rather than a search for "the delta," the range of days that, after smoothing, best fits the data was searched. More precisely, for each community, weights from a discrete triangle distribution were used to smooth the ultra-low sulfur diesel (ULSD) values. Then the ability of the smoothed index data to forecast the delivered price observed in the community was assessed. After iterating across a range of reasonable smoothing values, the range of days for which the smoothed values based on this range best fits the community data was selected. Finally, this process was repeated for each ice-free community.

Generating a smoothed ULSD value using the discrete triangle distribution requires the specification of both a centering value as well as a bandwidth. The centering value is the reported ULSD price that will serve as the center of the period across which to smooth. The centering value is also the ULSD reported price that will receive the greatest weight in the smoothing procedure. For example, if the centering value in a given community was 14, then for each delivery in the community, the ULSD reported price occurring two weeks prior to the data of delivery will be the centering delta and will have the largest effect on the smoothed value. The bandwidth specifies the number of ULSD reported prices that will be used in the smoothing. Only ULSD reported prices that are within the bandwidth will receive non-zero weight, with observations in the bandwidth receiving less weight as the distance from the centering value increases. For example, if the centering value was 14 and the bandwidth was 3, then a total of 7 observations would receive non-zero weight. The ULSD reported prices from the eleventh through seventeenth day prior to delivery will be used in the smoothing. In this example, the weights from the discrete triangle distribution would be $\left\{\frac{1}{16}, \frac{1}{8}, \frac{3}{16}, \frac{1}{4}, \frac{3}{16}, \frac{1}{8}, \frac{1}{16}\right\}$. The inner product of the vector of weights and the vector of ULSD reported prices in the bandwidth will be the smoothed ULSD value used in the algorithm.

The centering value and the bandwidth discussed above are tuning parameters that are learned from the data. In order to learn the centering value and bandwidth that bests fits the data for a given community, an algorithm was implemented. First, a smoothed ULSD value for each delivery using weights from the discrete triangle distribution, as discussed above is generated.

Then the data are randomly divided into two groups of roughly equal sizes. Each partition of the data consists of roughly half of the observations for the community. Each observation consists of a delivered price and the associated smoothed ULSD value. Next one of the two groups of data is randomly selected. The data in this group is used to estimate a simple linear regression of the form:

$$Delivered\ Price_i = \alpha + \beta\ Smoothed\ ULSD_i$$

The parameter estimates from the model (i.e., $\widehat{\alpha}$, $\widehat{\beta}$) are used to forecast the data in the group not used to estimate the model. Then, the root mean square prediction error is calculated. The data partitioning, model estimation, forecasting, and calculation of the root mean squared error are repeated several times yielding a vector with the elements of the vector being the root mean square prediction error for each iteration. Then the root mean squared prediction errors are averaged across runs. This process is repeated for every combination of centering value and bandwidth. The centering value and bandwidth that minimizes average root mean squared prediction error is selected.

The results reported in this work search across the range of centering values from one to thirty days, with bandwidth search occurring between one to fourteen days. Twenty-five replications of data partitioning, model estimation, forecasting, and calculation of the mean squared error were conducted.

Fixed Effects Least Squares Dummy Variable Estimator

The fixed effects estimator is a method for estimating an unobserved effects model. It involves a time-demeaning transformation that removes the unobserved fixed effect before the model is estimated (Wooldridge, 2006). It is assumed that there is endogeneity in the model due to correlation between the unobserved fixed effects and the independent variables. The fixed effects estimator helps control unobserved heterogeneity when the heterogeneity is both time-constant and correlated with the model's independent variables.

A general unobserved effects model is

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + a_i + u_{it}, t = 1,2, \ldots, T$$

In the notation y_{it}, *i* denotes the community and *t* denotes the time period. The unobserved, time-constant factors that affect y_{it} are captured in a_i. Here a_i is an unobserved community effect which represents all factors affecting delivered fuel price to a community that do not change over time. The idiosyncratic error, u_{it}, represents the time-varying unobserved factors that affect y_{it}. The fixed effect transformation involves first averaging the above equation over time for each *i*

$$\overline{y_i} = \beta_1 \overline{x_i} + \beta_2 \overline{x_i} + \cdots + \beta_k \overline{x_i} + a_i + \overline{u_i}$$

where $\overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{x_i} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, and $\overline{u_i} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. Next the averaged equation is subtracted from the initial equation yielding

$$y_{it} - \overline{y_i} = \beta_1(x_{it} - \overline{x_i}) + \beta_2(x_{it} - \overline{x_i}) + \cdots + \beta_k(x_{it} - \overline{x_i}) + u_{it} - \overline{u_i}, t = 1,2, \ldots, T$$

Each variable is time-demeaned by subtracting the variable average from each observation. By subtracting the averaged equation from the initial equation, the fixed effect a_i is removed. Any time-constant variables are also removed from the equation by time-demeaning. The model is then fit using pooled OLS regression on the time-demeaned variables. The time demeaned equation is

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it1} + \beta_2\ddot{x}_{it2} + \cdots + \beta_k\ddot{x}_{itk} + \ddot{u}_{it}, t = 1,2, \ldots, T$$

where $\ddot{y}_{it} = y_{it} - \overline{y}_i$ and so forth for $\ddot{x}_{it1}, \ddot{x}_{it2}, \ddot{x}_{itk}$, and $\ddot{u}_{it}$.

The least squares dummy variable (LSDV) estimator is pooled OLS with the inclusion of a dummy variable for each *i*. The LSDV estimator is mathematically equivalent to the fixed effect estimator in that the dummy variable regression yields the same parameter estimates, standard errors, and other major statistics as a regression on time-demeaned data (Schmidheiny, 2014). The fixed effect model assumes a_i is a parameter that can be estimated for each *i*. An intercept for each community can be estimated by adding a dummy variable for each cross-sectional observation that takes the value of one for community *i* and zero otherwise. Note that one community must serve as the base group to avoid the dummy variable trap. The overall model intercept serves as the intercept for the base group and the individual a_i are shifts in the intercept for the other communities (Dougherty, 2013).

Fuel Price Models

The population model for barge communities is

$$\ln(\text{Delivered Price}) = \beta_0 + \beta_1\ln(\text{Rbrent}) + a_2\text{ID}_2 + \cdots + a_n\text{ID}_n + u$$

where *ln(Delivered Price)* is the natural log of the observed price per gallon (2013$) of fuel at delivery, *ln(Rbrent)* is the natural log of inflation-adjusted smoothed price of Brent (2013$) on the estimated date of lift, and *ID* is a community-specific dummy variable. The term $\mathbf{a_i ID_i}$ represents the fixed effect on the dependent variable, *ln(Delivered Price)*, for community *i* when added to the intercept. The index for *ID* begins at 2 because one community is omitted as the base group to avoid the dummy variable trap. The intercept β_0 serves as the fixed effect for the base group. The coefficients on the community dummy variables shift the intercept yielding the fixed effects for other communities. The fixed effect encompasses any time-invariant factor affecting that community. For example, distance and the state of a community's mooring facilities would be absorbed in the community-specific fixed effect.

The population model for road and air communities is

$$\ln(\text{Delivered Price}) = \beta_0 + \beta_1 \ln(\text{Rbrent}) + \beta_2 \text{AC} + a_2 \text{ID}_2 \cdots + a_n \text{ID}_n + u$$

where *ln(Delivered Price)* is the natural log of the observed inflation-adjusted price per gallon (2013$) of fuel at delivery, *ln(Rbrent)* is the natural log of the inflation-adjusted price of Brent (2013$) on the date of lift, *AC* is the annual fuel consumption of a community's electric utility in gallons, and *ID* is a community-specific dummy variable. Like in the previous model, the term $a_i ID_i$ represents the fixed effect on the dependent variable, *ln(Delivered Price)*, for community *i* when added to the intercept.

There are two notable differences between the population model for the barge communities and the population model for the road and air communities. The first difference is that a smoothed Brent value is not used for the air and road models. For air and road communities, it is assumed that fuel is delivered on the same day of lift. Thus, the inflation-adjusted price of Brent on the date of delivery is used in the model. The second difference is the addition of a variable for annual fuel consumption. Annual consumption was added to the model for road and air communities for two reasons. First, the maximum quantity carried by an airplane or truck is much smaller than what a barge could carry for a single delivery. Second, an airplane or truck is likely to deliver its entire load to a single community as opposed to a barge which carries fuel for many communities on a single voyage. One would expect greater annual quantity to have a positive impact on the price of air deliveries because the greater the demand, the more trips required to meet that demand adding to the freight cost. On the other hand, one would expect greater annual quantity for road deliveries to have a negative impact on the price of road deliveries because a community buying in bulk may receive a discounted price.

It is not known how oil prices affect transport costs. In general, some research has suggested that for barge transport it matters little as a factor input in the cost of transport (Palo, 2012); however, it is possible that oil prices are correlated with other factors of production, and that the cost of refined product affects community-specific slope parameters. The log-log specification constrains the parameter on Brent price across communities by mode of transport, but allows community-specific differences in slope when transformed from log to level form. The model specification ensures that if community *i* has higher delivered fuel prices than community *j* at low oil prices then this ordering is preserved at higher oil prices as well.

All models were estimated using STATA software. The model errors were checked for heteroskedasticity and normality using the Breusch-Pagan and Shapiro-Wilks tests, respectively.

Price Projections

The estimated individual community equations were used to project delivered fuel prices using the forecasted Brent crude oil prices for 2015 to 2040 from the Energy Information Administration's Annual Energy Outlook (2015). Annual fuel consumption by a community's electric utility is an independent variable in both the air and road models. For this variable, 2013 values for annual fuel consumption were used.

3. Results

Ice-bound Barge Model

$$\ln\left(\widehat{\text{Delivered Price}}\right) = -1.54 + 0.64\ln(\text{Rbrent}) - 0.07\text{ID}_2 + \cdots - 0.02\text{ID}_{96}$$

$$\text{n} = 2{,}179 \quad \text{Adj. } R^2 = .8855 \quad \text{RMSE} = .1273$$

The ice-bound barge model has 2,179 observations. The model has an adjusted R^2 of .8855 and a root mean square error of .1273. Because both the dependent variable and *ln(Rbrent)* are natural logs, a 1% increase in real Brent price is associated with a 0.64% increase in delivered fuel price. The intercept term serves as the fixed effect for the base group, Akiachak. The coefficients on the *ID* dummy variables shift this intercept up or down. Since the dependent variable is a natural log, the proper interpretation of the coefficients for the *ID* dummy variables is the percentage change in the predicted value of the dependent variable. For example, ID_2 is the dummy variable for Akiak with a coefficient of -0.07. Thus, $\%\Delta \ln\left(\widehat{\text{Delivered Price}}\right) = [100 * (e^{\text{ID}_i} - 1)]$. Inserting Akiak's coefficient value into the equation yields $[100 * e^{-0.07} - 1)] = -6.76$. For Akiak, the intercept of the model is shifted down 6.76% relative to the base group community, Akiachak. See "Ice-bound barge results" in the accompanying spreadsheet for all parameter estimates and their p-values.

Ice-bound Barge Model		
Breusch-Pagan Test	$\chi^2(1) = 7.18$	Probability > $\chi^2 = 0.0074$
Shapiro-Wilk Test	$z = 9.887$	Probability > z = 0.000
Drop if standardized residual > |4|		10 observations deleted
Breusch-Pagan Test	$\chi^2(1) = 1.57$	Probability > $\chi^2 = 0.21$

For the Breusch-Pagan test for the ice-bound barge model, the null hypothesis of constant variance of the errors is rejected. Dropping 10 observations that are extreme outliers with standardized residuals greater than the absolute value of 4 corrects the heteroskedasticity. For the Shapiro-Wilk test, the null hypothesis of normal distribution of the errors is rejected. Because the errors are not normally distributed, the standard errors for the model are bootstrapped. A bootstrapped standard error is the standard deviation of repeated parameter estimates (Kennedy, 2003).

Ice-free Barge Model

$$\ln\left(\widehat{\text{Delivered Price}}\right) = -2.14 + 0.77\ln(\text{Rbrent}) - 0.17\text{ID}_2 + \cdots - 0.09\text{ID}_{33}$$

$$\text{n} = 1{,}723 \quad \text{Adj. } R^2 = .8943 \quad \text{RMSE} = .1442$$

The ice-free barge model has 1,723 observations. The model has an adjusted R^2 of .8943 and a root mean square error of .1442. For this model, a 1% increase in the real price of Brent is associated with a 0.77% increase in delivered fuel price. See "Ice-free barge results" in the accompanying spreadsheet for all parameter estimates and their p-values.

Ice-free Barge Model		
Breusch-Pagan Test	$\chi^2(1) = 467.22$	Probability > $\chi^2 = 0.000$
Shapiro-Wilk Test	$z = 12.721$	Probability > z = 0.000
Drop if standardized residual > |4|		10 observations deleted
Breusch-Pagan Test	$\chi^2(1) = 8.40$	Probability > $\chi^2 = 0.0037$

For the Breusch-Pagan test for the ice-free barge model, the null hypothesis of constant variance of the errors is rejected. Dropping 10 extreme outliers reduces but does not completely correct the heteroskedasticity. The presence of heteroskedasticity does not bias parameter estimates so model predictions are not impacted (Wooldridge, 2006). The null hypothesis of the Shapiro-Wilk test is rejected. The standard errors for the model are bootstrapped.

Air Model

$$\ln(\widehat{\text{Delivered Price}}) = -0.69 + 0.47\ln(\text{Rbrent}) + 0.0000029\text{AC} + 0.15\text{ID}_2 + \cdots - 0.39\text{ID}_{32}$$

$$\text{n} = 1{,}580 \quad \text{Adj. R}^2 = .7829 \quad \text{RMSE} = .1249$$

The air model has 1,580 observations. The model has an adjusted R^2 of .7829 and a root mean squared error of .1249. For this model, a 1% increase in the real price of Brent is associated with a 0.47% increase in delivered fuel price. Additionally, a 1% increase in annual fuel consumption by the electric utility is associated with a 0.00029% increase in delivered fuel price. See "Air results" in the accompanying spreadsheet for all parameter estimates and their p-values.

Air Model		
Breusch-Pagan Test	$\chi^2(1) = 158.26$	Probability > $\chi^2 = 0.000$
Shapiro-Wilk Test	$z = 11.181$	Probability > z = 0.000
Drop if standardized residual > |4|		9 observations deleted
Breusch-Pagan Test	$\chi^2(1) = 8.39$	Probability > $\chi^2 = 0.0038$

For the Breusch-Pagan test for the air model, the null hypothesis of constant variance of the errors is rejected. Dropping 9 extreme outliers reduces but does not completely correct the heteroskedasticity. The null hypothesis of the Shapiro-Wilk test is rejected. The standard errors for the model are bootstrapped.

Road Model

$$\ln(\widehat{\text{Delivered Price}}) = -2.19 + 0.77\ln(\text{Rbrent}) - 0.000000293\text{AC} - 0.15\text{ID}_2 + \cdots + 0.09\text{ID}_{14}$$

$$\text{n} = 1{,}341 \quad \text{Adj. R}^2 = .9253 \quad \text{RMSE} = .1015$$

The road model has 1,341 observations. The model has an adjusted R^2 of .9253 and a root mean squared error of .1015. Here a 1% increase in the real price of Brent is associated with a 0.77%

increase in delivered fuel price. Also, a 1% increase in annual fuel consumption by an electric utility is associated with a 0.0000293% decrease in delivered fuel price. See "Road results" in the accompanying spreadsheet for all parameter estimates and their p-values.

Road Model		
Breusch-Pagan Test	$\chi^2(1) = 113.02$	Probability > $\chi^2 = 0.000$
Shapiro-Wilk Test	$z = 8.516$	Probability > z = 0.000
Drop if standardized residual > |4|		0 observations deleted

For the Breusch-Pagan test for the road model, the null hypothesis of constant variance of the errors is rejected. Dropping outliers does not improve the heteroskedasticity in this case, so no outliers are dropped. As mentioned before, heteroskedastiticy does not bias parameter estimates, so model predications are not affected. The null hypothesis of the Shapiro-Wilk test is rejected. The standard errors for the model are bootstrapped.

Price Projections

The results of the fuel price projections for forecasted Brent prices from 2015 to 2040 for the ice-bound barge model, ice-free barge model, air model, and road model are available in the accompanying spreadsheet in sheets labeled "Ice-bound barge projections," "Ice-free barge projections," "Air projections," and "Road projections," respectively.

4. Discussion

The goal of this research was to build community-specific models that can be used to predict the delivered price of fuel to rural communities throughout Alaska. The LSDV estimator is an appropriate method because it provides community-specific fixed effects. Time-constant variables for which there is no good source of publicly-available data are absorbed into the community-specific fixed effect. For example, there is a lack of publicly-available information on fuel storage and mooring facilities in rural communities. Fuel storage ownership and capacity as well as the state of a community's mooring facilities likely impact the delivered price of fuel. However, even if this information was available, these are time-constant variables. Thus, they would be removed from the model when the variables are time-demeaned. Using the LSDV estimator, they are absorbed into the community-specific fixed effects. The fixed effect term represents all time-constant omitted variables that have been controlled for through fixed effects transformation.

A logical consequence of the log-log model framework is that as oil prices approach zero, then projected delivered refined product prices also tend toward zero. In the limit delivered prices cannot be zero, even assuming costless crude, given positive costs of transportation and delivery logistics. Oil prices during the time period covering the dataset are large enough to prevent this dynamic.

The model specification over the relevant range of prices allows for comparatively modest increases in the differences in delivered prices across communities within a mode of transport as oil prices rise. Had the model been specified in level-level form, rather than in log-log form, each community would have its own intercept but the slope on the Brent coefficient would not vary within mode of transport. However, specifying the model in level-level form worsens heteroskedasticity in the errors. This suggests that delivered price differences might indeed increase with fuel prices.

A primary innovation of this work is its reliance upon invoice data, which allows for improved time resolution in the largest determinant of delivered price, i.e., the crude oil price at the refinery price at the time of lift. Nevertheless time resolution remains imperfect. First, invoice inspection does not reveal whether fuel was pulled from storage. This 'smears' the relationship between predicted and actual lift dates. If fuel spends a month in storage before it is delivered, the estimate of the delta between lift date and delivery date will be inaccurate; the month in storage is not accounted for. Second, invoices typically do not reveal where fuel is sourced. It could come from Alaska, Seattle, or Asia, among other places. This is a source of measurement error because distance matters in the actual time delta between lift and delivery.

Both of these shortcomings could be resolved if fuel distributors listed the date of fuel lift from the refinery on the invoice. This applies even to fuel placed in storage tanks before delivery. Knowing the date on which stored fuel was lifted would greatly increase the accuracy of the model. Additionally, knowing where fuel was sourced loses its importance when the lift date is known. The delta between lift and delivery would not need to be estimated, so it would no longer be necessary to know the distance the fuel traveled. It may not be an accident that the road delivery model produces the tightest correlations in the data, as it seems unlikely that communities on the road system often take delivery of fuel that has been first put into storage.

5. Conclusion

This project used a unique data set comprised of rural utility fuel invoices collected from the RCA. Data was separated by mode of delivery to account for the different price processes governing each mode of transportation. The delta between lift and delivery was estimated for barge communities, and assumed to be contemporaneous with delivery date for air and road communities. The LSDV estimator was used to estimate community-specific rural fuel models. These models can be used to project delivered fuel price to rural communities. The projections can be used to assess the economic feasibility of proposed renewable energy projects for the Renewable Energy Fund. The models could be improved in the future if fuel distributors listed the lift date of delivered fuel on the invoice.

The database should be updated annually to keep the task manageable. Data collection is the most time-intensive component of this project. It should be possible for an employee working fulltime to update the database in a summer. All new fuel invoices should be added to the

database. The models should be re-specified annually or biennially. Periodically, the functional form of the model should be reassessed as more data is added to the database. The estimates of the community-specific fixed effects will improve as additional fuel invoices are added to the dataset.

Acknowledgments
The authors thank Camilla Kennedy, Samuel Tappen, Benjamin Wilke, Jonathan Quinones, Henry Arend, and Chandler Kemp for collecting data. Additionally, the authors thank Jeremy Vandermeer for writing the MatLab code for the ice-bound barge price weighting, and Henry Arend for writing instructions for updating the database.

References

Alaska Department of Revenue Tax Division. (2015). *Crude oil and natural gas prices* [webpage]. Retrieved from http://tax.alaska.gov/programs/oil/dailyoil/dailyoil.aspx

Alaska Energy Authority. (2015). *Renewable energy fund* [webpage]. Retrieved from http://www.akenergyauthority.org/Programs/RenewableEnergyFund

Allison, Paul D. (2009). "Missing Data." in *The SAGE Handbook of Quantitative Methods in Psychology*. Roger E. Millsap and Alberto Maydeu-Olivares (Eds). Thousand Oaks, CA: Sage Publications Inc.

Dougherty, C. (2013). *Fixed effects regression: LSDV method* [PowerPoint slides]. Retrieved from http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB8QFjAAahUKEwjbgcDV5ZLGAhVIO4gKHbARAGw&url=http%3A%2F%2Flearningresources.lse.ac.uk%2F140%2F3%2FChapter%252014%2520fixed%2520effects%2520regressions%2520least%2520square%2520dummy%2520variable%2520approach%2520%28EC220%29.pps&ei=4Vd_VZuNDMj2oASwo4DgBg&usg=AFQjCNGMa2BNsk0UYqClPdKlDGM-5mTzjQ&bvm=bv.95515949,d.cGU

Energy Information Administration (2015). Annual energy outlook. Retrieved from http://www.eia.gov/forecasts/aeo/

Kennedy. P. (2003). *A guide to econometrics* (5th ed.). Cambridge, Massachusetts: The MIT Press.

National Oceanic and Atmospheric Administration. (2012) *Distance between United States ports* [PDF document]. Retrieved from http://www.nauticalcharts.noaa.gov/nsd/distances-ports/distances.pdf.

Palo, G. (2012). On maritime transport costs, evolution, and forecasts. *Ship and Science Technology*, *10*(5), 19-31.

Schmidheiny, K. (2014). *Panel data: Fixed and random effects* [PDF document]. Retrieved from http://kurt.schmidheiny.name/teaching/panel.pdf

Syzmoniak, N., Fay,. G., Villalobos-Melendez, A., Charon, J., & Smith, M. (2010). *Components of Alaska fuel costs: An analysis of the market factors and characteristics that influence rural fuel prices*. University of Alaska Anchorage Institute of Social and Economic Research.

United States Bureau of Labor Statistics. (2015). *CPI detailed report* [PDF document]. Retrieved from http://www.bls.gov/cpi/cpid1504.pdf

Wooldridge, J. (2006). *Introductory econometrics: A modern approach* (3rd ed.). Mason, Ohio: Thompson South-Western

Appendix 1: Updating Database and Models

Updating Database

1. Connect to the Regulatory Commission website by enter the URL rca.alaska.gov
2. Mouse over the RCA Library tab and click Advanced Search
3. Enter the name of the community you are searching for enclosed in double quotes
4. Use the start and end dates per the following
    a. Start: 1/1/14
    b. End: whatever the current date is
5. Select the check box labeled "Limit search to the selected filing types"
6. Taking great care to select multiple types, select:
    a. R&F – PCE (Non-Regulated)
    b. R&F – PCE Annual Report (Non-Regulated)-52.660(c)
    c. R&F – PCE Fuel & Purchased Power Cost Request/Report (Non-regulated)-52.640(b)&(f)(2)
    d. R&F – PCE (Non-regulated) Compliance Filing
    e. R&F – PCE (Non-regulated) Supplemental Filing
7. Enter the search and a list of 25 results will be shown from the drop down menu on the side select 200 results per page
8. Using the browsers search function find any entry that contains the keyword fuel. (most major browsers will highlight the different instances of fuel and you can manually scroll through them and open the PDFs as you go along)
9. By opening multiple PDFs at once efficiency can be increased greatly however be careful not to open too many tabs as it will crash your computer
10. Looking through the different entries some will be found to be more useful than others also some will be the PCE fuel forms while others will show the actual receipts.
11. After all of the invoices have been read through, the last entry date regardless of what type of entry will be used as the new end date and these instructions can be repeated from step 4 until the entries reaching 1/1/14 have been reached.
12. Additional invoices not in the online database can be found through an in-person search at the Regulatory Commission of Alaska.
13. For each invoice, the community name, delivered fuel price per gallon, delivery date, payment terms, fuel type, quantity, seller, the electric utility, and mode of transportation should be recorded in a row in the master spreadsheet.
14. Anomalous invoices should be coded in the column of the spreadsheet labeled "Invoice Issues" as follows:

1	No proof of original transaction (i.e. invoice or receipt) information on PCE form or cost worksheet only
2	Fuel Purchased from Local Store/School District
3	Original Invoice no longer in RCA online, (example- unlike the case of an entry coded "1" where we typically know: the seller, terms, type of fuel etc. because at one point info was taken from an actual invoice)
4	Invoice Illegible/difficult to read
5	Possible human error
6	Other, drastic issues please check comments, may need to be dropped

7 Multi-community fuel shipment/multiple customers
8 Price cap per fuel contract (where market price is present it is noted in the comments section)
9 Fuel resale with Markup

Updating Ice-bound Barge Model

1. Copy all observations for ice-bound barge deliveries from master spreadsheet into new spreadsheet.
2. Sort data by "AVECdelta" and pull out all entries coded "1" which are colored purple. Set them aside in a different spreadsheet. These entries have known lift dates and not need smoothed prices.
3. Use "AVEC ice-bound delta data" dataset to estimate model to predict the delta between lift and delivery in STATA software where the natural log of the delta between lift and delivery is the dependent variable and distance and a dummy variable for Yukon River community are the independent variables.
4. Use this model with your ice-bound barge community data to generate a predicted delta, standard error, and confidence interval associated with the predicted value for each observation in the ice-bound barge dataset.
5. Copy these values into a spreadsheet plus the date associated with each observation
6. Obtain Brent prices from Alaska Department of Revenue Tax Department and add them plus their associated dates to the existing Brent prices already collected.
7. Use MatLab software to generate a weighted Brent value associated with each observation.
8. Pull weighted Brent values back into ice-bound barge data spreadsheet.
9. Copy "AVECdelta" observations back into the ice-bound barge spreadsheet.
10. Adjust delivered fuel price per gallon and weighted Brent price for inflation using the U.S. Bureau of Labor Statistics Historical Consumer Price Index for All Urban Customers (CPI-U).
11. Assign a unique ID variable for each community.
12. Remove any observation associated with December, January, or February delivery dates since barge deliveries cannot be made in the winter.
13. Estimate model using LSDV estimator in STATA software where the natural log of the delivered price per gallon is the dependent variable and the natural log of inflation-adjusted, weighted Brent and community-specific dummy variables are the independent variables.

Updating Ice-free Barge Model

1. Copy all observations for ice-free barge deliveries from of master spreadsheet into new spreadsheet.
2. Collapse consecutive entries for a community with identical prices into first entry with that price.
3. For a given choice of centering value and bandwidth, generate a smoothed ULSD value for each delivery using weights from the discrete triangle distribution in R software.
4. Randomly divide the data into two groups of roughly equal sizes. Each partition of the data will consist of roughly half of the observations for the community. Each observation consists of a delivered price and the associated smoothed ULSD value generated in step 1.
5. Randomly pick one of the groups of data. Use the data in this group to estimate a simple linear regression of the form:
    i. $Delivered\ Price_i = \alpha + \beta\ Smoothed\ ULSD_i$
6. Use the estimates from the model in step 3 (i.e., $\hat{\alpha}$, $\widehat{\beta}$) to forecast the data in the group not used to estimate the model in step 3.
7. Calculate the root mean square prediction error
8. Repeat steps 2 through 5 twenty-five times. This yields a vector with the elements of the vector being the root mean square prediction error for each iteration. Average the root mean squared prediction error across runs
9. Repeat steps 1 through 6 for every combination of centering value and bandwidth
10. Select the centering value and bandwidth that minimizes average root mean squared prediction error
11. Transfer smoothed Brent values back to the spreadsheet. Adjust both the delivered price per gallon and the smoothed Brent prices for inflation
12. Assign each community a unique ID number
13. Estimate model using LSDV estimator in STATA software where the natural log of the delivered price per gallon is the dependent variable and the natural log of inflation-adjusted, smoothed Brent and community-specific dummy variables are the independent variables.

Updating Air Model

1. Copy all observations for air communities from of master spreadsheet into new spreadsheet.
2. Obtain Brent prices from Alaska Department of Revenue Tax Department
3. Pull the Brent price associated with the delivery date on the invoice into spreadsheet for each observation
4. Obtain utility annual fuel consumption from PCE Statistical Reports.
5. Pull utility annual fuel consumption associated with the year from the delivery date on the invoice into spreadsheet for each observation.
6. Adjust delivered fuel price per gallon and Brent price for inflation using the U.S. Bureau of Labor Statistics Historical Consumer Price Index for All Urban Customers (CPI-U).
7. Assign a unique ID variable for each community.
8. Estimate model using LSDV estimator in STATA software where the natural log of the delivered price per gallon is the dependent variable and the natural log of inflation-adjusted Brent, utility annual fuel consumption, and community-specific dummy variables are the independent variables.

Updating Road Model

1. Copy all observations for road deliveries from of master spreadsheet into new spreadsheet.
2. Collapse consecutive entries for a community with identical prices into first entry with that price.
3. Obtain Brent prices from Alaska Department of Revenue Tax Department
4. Pull the Brent price associated with the delivery date on the invoice into spreadsheet for each observation
5. Obtain utility annual fuel consumption from PCE Statistical Reports
6. Pull utility annual fuel consumption associated with the year from the delivery date on the invoice into spreadsheet for each observation
7. Adjust delivered fuel price per gallon and Brent price for inflation using the U.S. Bureau of Labor Statistics Historical Consumer Price Index for All Urban Customers (CPI-U).
8. Assign a unique ID variable for each community
9. Estimate model using LSDV estimator in STATA software where the natural log of the delivered price per gallon is the dependent variable and the natural log of inflation-adjusted Brent, utility annual fuel consumption, and community-specific dummy variables are the independent variables.

Appendix 2: Code for Models

Ice-bound Barge Model

Ice-bound barge delta code for STATA

import excel "File path", sheet("AVEC ice-bound delta data") firstrow

gen lnlag=log(Lag)

bootstrap, reps(1000) seed(1): reg lnlag Distance YR

import excel “File path”, sheet("Ice-bound barge data") firstrow clear

predict yhat

predict stdp, stdp

generate tmult=invttail(505, .025)

generate lowerPI=yhat-tmult*stdp

generate upperPI=yhat+tmult*stdp

*This code yields the predicted values, standard errors, and lower and upper bound for the prediction intervals. The antilog of each value should be taken. Place these results in a spreadsheet in the following order: a unique ID for each observation, the predicted value, the lower bound of the prediction interval, the upper bound of the prediction interval, the standard error of the prediction, and the date associated with the observation. Your spreadsheet should have six columns. Save this spreadsheet in your MatLab directory as ‘barge’. Create a separate spreadsheet with dates in the first column and associated Brent crude prices in the second column. Save this spreadsheet in your MatLab directory as ‘dates’.

The date format must first be changed. Then the program for smoothing the Brent prices can be run. Make sure none of the upper or lower bounds in in your prediction intervals are outside of your Brent price dataset. For example if your last date in your Brent price dataset is December 31, 2013, make sure none of the bounds of any confidence interval is beyond December 31, 2013 for which you do not have price data. If this happens you will get an error message. This is why you must have a unique ID number for each observation. The unique ID will allow you to easily identify any problem observation.

Code for changing date format program for MatLab

```
[ndata,text,alldata]=xlsread('barge.xlsx',6,'A1:F25'); %6 is the sheet
number, 'A1:F25' is the data range

B=zeros(length(text(:,6))-1,1); %6 is the column with the date
```

```
for i=2:length(text(:,6))
    B(i-1)=datenum(text{i,6});
end

A = [ndata,B]; %attaches new B date vector to numerical data in one matrix

[ndata_money,text_money,alldata_money]=xlsread('dates.xlsx');

B_money=zeros(length(text_money(:,1))-1,1);

for i=2:length(text_money(:,1))
    B_money(i-1)=datenum(text_money{i,1});
end

Prices = [B_money,ndata_money];
```

Code for weighted prices program for MatLab

```
B=cell(length(A(:,1)),1);
for i=1:length(B)
    B{i}.prob = pdf('normal',(A(i,3)):(A(i,4)),A(i,2),A(i,5));    %returns
probabilities for each day
    B{i}.delta = A(i,3):A(i,4); %difference between upper and lower PI bounds
    B{i}.del_date = A(i,6); %assuming a 6th column includes delivery date

    B{i}.date = B{i}.del_date -B{i}.delta; %subtracts numbers w/in the PI
from the delivery date

    for j=1:length(B{i}.date)
        [a,ind]=histc(B{i}.date(j),Prices(:,1)); %Prices is separate matrix,
1st column date, 2nd price
        B{i}.Prices(j) = Prices(ind,2);
    end

    B{i}.Weighted_prices = B{i}.Prices .* B{i}.prob; %multiplies prices by
probabilities
    B{i}.Sum_price = sum(B{i}.Weighted_prices); %sums weighted prices
end

xlswrite('barge.xlsx',Sum_prices,'Sum')    % write back to excel sheet on
% sheet named 'Sum'
```

* In MatLab first run code for changing format program. Then run code for weighted prices program. The summed weighted prices are written to a sheet named “Sum” in the barge excel sheet. Copy and paste these prices into your Ice-bound Barge dataset. Adjust these prices for inflation. Save the spreadsheet. These prices will be used in your regression.

Ice-bound barge model program for STATA

```
import excel "File path", sheet("Ice-bound barge data") firstrow

gen lRprice=log(Rprice)
```

```
gen lRbrent=log(Rbrent)

reg lRprice lRbrent i.ID

predict rs, rstandard

estat hettest

swilk rs

drop if rs > 4

drop if rs < -4

estat hettest

swilk rs

reg lRprice lRbrent i.ID, vce( bootstrap, reps (1000) seed(1) )
```

Ice-free Barge Model

Ice-free barge delta code for R

```
##################################################

##### "Smoothed Regression" Approach##########

##################################################

#The idea is to do the following:

#  1.)Pick a lag to center the weighting function on, pick the weighting function (i.e., the kernel),

#        and pick the number lags allowed to enter the weighting (i.e., the bandwidth)

#  2.)For a given delivery, apply the weights to the selected lagged INDEX values.  This yields a weighed INDEX value

#  3.)Construct the ordered triple of (Date of Delivery, Delivery Price, weighted INDEX value)

#  4.)Repeat for all deliveries

#  5.)Perform 2-k crossvalidation several times regressing delivered price on the weighted INDEX vector

#  6.)Iterate across bandwith and centering lag, keeping the average CV loss for each bandwith/centering lag combination

#  7.)Pick model (i.e., bandwith/centering lag) that minimizes CV loss

#

#Impose a constraint such that the centering lag/bandwidth combo doesn't allow future values to enter in the weighting scheme

##################################################

##################################################

#Directly calculate the weights for a discrete triangle distribution


DT.3<-c(1,2,1)*1/sum(c(1,2,1))

DT.5<-c(1,2,3,2,1)*1/sum(c(1,2,3,2,1))

DT.7<-c(1,2,3,4,3,2,1)*1/sum(c(1,2,3,4,3,2,1))

DT.9<-c(1,2,3,4,5,4,3,2,1)*1/sum(c(1,2,3,4,5,4,3,2,1))

DT.11<-c(1,2,3,4,5,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,5,4,3,2,1))
```

```
DT.13<-c(1,2,3,4,5,6,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,6,5,4,3,2,1))

DT.15<-c(1,2,3,4,5,6,7,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,7,6,5,4,3,2,1))

DT.17<-c(1,2,3,4,5,6,7,8,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,8,7,6,5,4,3,2,1))

DT.19<-c(1,2,3,4,5,6,7,8,9,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,10,9,8,7,6,5,4,3,2,1))

DT.21<-
c(1,2,3,4,5,6,7,8,9,10,11,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,10,11,10,9,8,7,6,5,4,3,2,
1))

DT.23<-
c(1,2,3,4,5,6,7,8,9,10,11,12,11,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,10,11,12,11,10,9,8
,7,6,5,4,3,2,1))

DT.25<-
c(1,2,3,4,5,6,7,8,9,10,11,12,13,12,11,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,10,11,12,13,
12,11,10,9,8,7,6,5,4,3,2,1))

DT.27<-
c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,13,12,11,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,10,11,
12,13,14,13,12,11,10,9,8,7,6,5,4,3,2,1))

DT.29<-
c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,14,13,12,11,10,9,8,7,6,5,4,3,2,1)*1/sum(c(1,2,3,4,5,6,7,8,9,
10,11,12,13,14,15,14,13,12,11,10,9,8,7,6,5,4,3,2,1))


DT.weight<-function(num.lag){

  if (num.lag==3){

    return(DT.3)

  }  else if (num.lag==5){return(DT.5)

  }  else if (num.lag==7){return(DT.7)

  }  else if (num.lag==9){return(DT.9)

  }  else if (num.lag==11){return(DT.11)

  }  else if (num.lag==13){return(DT.13)

  }  else if (num.lag==15){return(DT.15)

  }  else if (num.lag==17){return(DT.17)

  }  else if (num.lag==19){return(DT.19)
```

```
  }  else if (num.lag==21){return(DT.21)

  }  else if (num.lag==23){return(DT.23)

  }  else if (num.lag==25){return(DT.25)

  }  else if (num.lag==27){return(DT.27)

  }  else if (num.lag==29){return(DT.29)

  }else {return(cat("Hmm, I thought we were constraining to a bandwidth of 1 to 14?"))

  }

}

################################################

################################################

##Develop a function that, for each delivery, finds the center lag and appropriate number of leading and trailing lags and then applies the weighting function

################################################

#Delivery.date is a single date, reference data is a matrix with 1st column vector of dates and second column vector of prices, center.lag and num.lag are scalars

smoothed.DT<-function(delivery.data,reference.data,center.lag,num.lag){

  #Find the position of the center lag in the data

  index<-ifelse(sum(delivery.data-
reference.data[,1]>=center.lag)==0,"NA",tail(which(delivery.data-
reference.data[,1]>=center.lag),n=1))

  #Extract the index data associated with the supplied center lag and number of lags

  index.vector<-(seq(1:num.lag)-seq(1:num.lag)[(length(seq(1:num.lag))+1)/2])+index

  #Grab the prices that are associated with the indices

  data.for.smoothing<-reference.data[index.vector,2]

  #Now smooth

  smoothed.value<-sum(data.for.smoothing*DT.weight(num.lag))

  #Return the smoothed value

  return(smoothed.value)

}
```

```
#reference and community data are a matrix with the 1st column vector being dates and second column vector being prices

#returns a vector of the smoothed index prices, has the same length as the community data

smoothed.data<-function(community.data,reference.data,center.lag,num.lag){

  smoothed.vec<-rep(NA,length=dim(community.data)[1])

  for (i in 1:dim(community.data)[1]){

    smoothed.vec[i]<-smoothed.DT(community.data[i,1], reference.data, center.lag, num.lag)

    }

  return(smoothed.vec)

  }

#community.data is a two column matrix as usual, smoothed data is a vector with length equal to number of rows in community.data, user.K is the number of folds

#to use in cross-validation, user.R is a scalar indicating number of cv repetitions

#returns the average root mean squared prediction error across repetitions

cv.rmspe<-function(community.data,smoothed.vec,user.K,user.R){

  dum.data<-as.data.frame(cbind(community.data[,2],smoothed.vec))

  names(dum.data)<-c("communitydata","smoothedvec")

  a<-lm(communitydata~smoothedvec, data=dum.data)

  print(summary(a))

  b<-cvLm(a,K=user.K,R=user.R,cost=rtmspe)

  return(b)

}

#Wrapper function to automate the search across the center lag and bandwidth spaces

cv.smoothed.data<-function(community.data, reference.data, min.center.lag, max.center.lag, min.num.lag, max.num.lag, num.folds, num.reps){

  cv.matrix<-as.data.frame(matrix(NA, ncol=3,nrow=(((max.num.lag-min.num.lag)/2)+1)*(((max.center.lag-min.center.lag)/2)+1)))

  h<-0

  for (i in min.center.lag:max.center.lag){
```

```
    j<-min.num.lag
    while (j<=max.num.lag){
      h<-h+1
      smooth<-smoothed.data(community.data,reference.data,center.lag=i,num.lag=j)
      cv.out<-cv.rmspe(community.data,smooth, user.K=num.folds, user.R=num.reps)
      cv.matrix[h,1]<-i
      cv.matrix[h,2]<-j
      cv.matrix[h,3]<-cv.out$cv
      if ((((j-1)/2))>i)
      {
        cv.matrix[h,3]<-9999999999
      }
      j<-j+2
    }
  }
  names(cv.matrix)<-c("Centering Lag", "Num Lags Used in Weight", "Average RMSPE")
  return(cv.matrix)
}
```

Ice-free barge code for STATA

```
import excel "File path", sheet("Ice-free barge data") firstrow

gen lRprice=log(Rprice)

gen lRbrent=log(Rbrent)

reg lRprice lRbrent i.ID

predict rs, rstandard

estat hettest

swilk rs
```

```
drop if rs > 4

drop if rs < -4

estat hettest

swilk rs

reg lRprice lRbrent i.ID, vce( bootstrap, reps (1000) seed(1) )
```

Air Model

Air model code for STATA

```
import excel "File path", sheet("Air data") firstrow

gen lRprice=log(Rprice)

gen lRbrent=log(Rbrent)

reg lRprice lRbrent AC i.ID

predict rs, rstandard

estat hettest

swilk rs

drop if rs > 4

drop if rs < -4

estat hettest

swilk rs

reg lRprice lRbrent AC i.ID, vce( bootstrap, reps (1000) seed(1) )
```

Road Model

Road model code for STATA

```
import excel "File path", sheet("Road data") firstrow

gen lRprice=log(Rprice)

gen lRbrent=log(Rbrent)

reg lRprice lRbrent AC i.ID

predict rs, rstandard

estat hettest

swilk rs

drop if rs > 4

drop if rs < -4

estat hettest

swilk rs

reg lRprice lRbrent AC i.ID, vce( bootstrap, reps (1000) seed(1) )
```