r

pennsylvania
DEPARTMENT OF ENVIRONMENTAL PROTECTION

Office of Oil and Gas Management

DEP INFRASTRUCTURE INVESTMENT AND JOBS ACT (IIJA) FORMULA GRANT DOCUMENTATION

Rev Date: December 22, 2021

pennsylvania
DEPARTMENT OF ENVIRONMENTAL PROTECTION

Table of Contents

1.0 Abstract/Introduction

Federal legislation providing significant stimulus funding to plug orphan and abandoned oil and gas wells nationally was introduced on April 8, 2021 in the United States House of Representatives (U.S. House) (H.R. 2415). On May 26, House Natural Resources Committee voted out H.R. 2415 for consideration by the full House. Senator Ben Ray Lujan (D-NM) sponsored companion legislation in the Senate on April 12, 2021 (S. 1076). Such funding was also announced as part of President Biden's American Jobs Plan in Pittsburgh on March 31, 2021, increasing awareness of the issue, and indicating Executive Branch support for the concept. In a letter dated May 25, 2021, the Interstate Oil and Gas Compact Commission (IOGCC) formally supported S. 1076. The letter was signed by IOGCC member state Governors, including Governor Wolf. On August 10, 2021, the U.S. Senate passed the 2021 Infrastructure Investment and Jobs Act (IIJA) bill by a vote of 69-30. This $1 trillion infrastructure plan includes funding for plugging abandoned wells, with the language for these specific provisions mirroring the REGROW Act. The U.S. House passed the legislation on November 5, 2021 and it was signed into law by President Biden on November 15, 2021.

This report documents the methods used to derive the three criteria required for the Notice of Intent (NOI) to Apply for Formula Grant Funding. The three required factors on the NOI and explained in this report include:

- Job Losses
- Documented Orphaned Wells (on state or private land)
- Projected Costs

2.0 Job Losses

The first portion of the NOI requires the job losses in the oil and gas industry from the period beginning on March 1, 2020 and ending on November 15, 2021. Per IOGCC's recommendation, DEP coordinated with Mr. Edward Legge, Director of the Center for Workforce Information and Analysis at the Department of Labor and Industry. Mr. Legge is Pennsylvania's Labor Market Information contact and can be contacted at elegge@pa.gov or 717.787.8646.

On November 23, 2021, Mr. Legge indicated that the November 2021 data requested from the NOI will not be available until February 2022. He was able to provide job loss numbers from March 1, 2020 through June 30, 2021 for the Oil and Gas Extraction Industry (NAICS 2011) as well as the categories for Drilling Oil and Gas Wells (NAICS 213111) and Support Activities for Oil and Gas Operations (NAICS 213112) (Table 1).

There were 4,569 jobs in the Oil and Gas Extraction Industry (NAICS 211) in PA in March of 2020 compared to 3,921 in June of 2021. This is a decline of 648 jobs or 14.2%.

There were 1,318 jobs in the Drilling Oil and Gas Wells Industry (NAICS 213111) in PA in March of 2020 compared to 988 in June of 2021. This is a decline of 330 jobs or 25.0%. There were 8,792 jobs in the Support Activities for Oil and Gas Operation (NAICS 213112) in PA in March of 2020 compared to 6,749 in June of 2021. This is a decline of 2,043 jobs or 23.2%. The totals for the oil

and gas 213 categories is 10,110 jobs in March 2020 and 7,737 jobs estimated for November 15, 2021. This is a decline of 2,373 jobs or 23.5%.

The data developed in Pennsylvania during 2020 and 2021 shows continued slight declines in 211 and 213111 and a leveling off in 213112. Based on the trend over the last year plus, Pennsylvania estimated the job numbers on November 15, 2021 as being nearly identical to the June 30, 2021 numbers, rendering any differences inconsequential.

NAICS	*Jobs – March 1, 2020*	*Jobs – November 15, 2021*	*Job Loss*	*% Change*
211	4,569	3,921	648	-14.2%
213 (total)	10,110	7,737	2,373	-23.5%
TOTAL	**14,679**	**11,658**	**3,021**	**-20.6%**

Table 1. Job losses in oil and gas sector in Pennsylvania between March 1, 2020 and November 15, 2021.

3.0 Documented Orphaned Wells

IIJA defines "Orphaned" well in the context of the state grant program as *the name given the term by the applicable state or, if that state uses different terminology, has the meaning given another term used by the state to describe a well eligible for plugging, remediation, and reclamation by the state.*

In the 2012 Oil and Gas Act, Pennsylvania has definitions for both *orphan* and *abandoned* wells:

- *Orphan:* a well abandoned prior to April 18, 1985, that has not been affected or operated by the present owner or operator and from which the present owner, operator or lessee has received no economic benefit other than as a landowner or recipient of a royalty interest from the well.
- *Abandoned:* Any of the following: (1) A well: (i) that has not been used to produce, extract or inject any gas, petroleum or other liquid within the preceding 12 months; (ii) for which equipment necessary for production, extraction or injection has been removed; or (iii) considered dry and not equipped for production within 60 days after drilling, redrilling or deepening. (2) The term does not include wells granted inactive status.

Both of the classes of wells above are eligible for plugging, remediation, and reclamation by the commonwealth, although Pennsylvania does not typically advance well decommissioning contracts until all responsible party searchers are exhausted or in the event that the well is contributing to immediate threats to public health and safety. In the latter case, plugging may commence and cost recovery may be pursued if a responsible party is later identified.

For the purposes of the NOI, five classes of wells have been categorized as "Orphaned" in a manner consistent with the IIJA definition – note that the group nomenclature for the first 4 classes has internal significance at DEP:

- *Group 1:* wells associated with operators who are bankrupt, deceased, out-of-business, or no longer conducting business with the agency.

- *Group 2:* wells identified by operators in 2020 as abandoned in accordance with the well identification provisions of Act 13, Section 3213.
- *Group 5:* wells identified by DEP field staff as abandoned.
- *Group 6:* 90000-series wells classified as "Abandoned" – pre-Act/legacy wells generally reported in a historical oil and gas publication (William McGlade Study) that were field verified and never registered.
- *DEP Orphan and DEP Abandoned Wells:* wells meeting the statutory definition of *Orphan* or *Abandoned* for which all efforts to identify a responsible party have been exhausted and no viable owner or operator exists.

Excluding wells known to exist on federal lands, the total number of wells on state or private lands from the above five groups is 26,908. Note that 9,541 of these wells do not have locations documented electronically in DEP records.

4.0 Projected Costs

DEP has a long history of developing and executing plugging contracts in the state that dates back to 1989. The agency has compiled records for over 3,000 wells that have been decommissioned over this history. Historical plugging cost and other contract data are useful for predicting future plugging costs. To do so, the agency used a standard machine learning approach to develop a Generalized Additive Model (GAM).

GAMs represent a supervised learning technique. Like multiple linear regression, GAMs may be used to fit functions for a number of individual predictor variables. The added power associated with GAMs is that, unlike multiple linear regression, the functions may be non-linear in character. GAMs also are compatible with both continuous and categorical variables – both for the predictor (classification) and the dependent variable (James et al., 2017).

James et al. (2017) provide more detail about the mathematical framework associated with GAMs. The overall objective is similar to multiple linear regression – fit a function to one or more predictor variables that allows the response variable to be predicted in a manner that reduces the error as much as possible. In other words, more precisely fitted models have less residual error and the predictions (estimated values) are closer to the observations (actual values) of the dependent variables. In this case, DEP is interested in developing a model that predicts expected per-well plugging costs for future contracts.

Conceptualizing a GAM is easy for up to two predictor variables. The figure below shows a response variable (vertical axis) predicted as a function of two predictor variables for two multivariate models (Figure 1). The first conceptual model shows a multiple linear regression fit – note the prediction surface is planar, i.e., linear. The vertical offset between the actual data points (red spheres) and the model fit is referred to as the residual error and it can be reported as the mean square error (MSE) by squaring each error measurement (RSS), summing the squares, and dividing by the total number of cases. By taking the square root of MSE, the Root MSE (RMSE) is derived and it represents on average how close predicted values are to observed values for the

model or learning method. The second conceptual model has been fitted using a smoother function as the basis function in a GAM, in this case splines are used – note the raw data used to develop the modeled surface are not shown, but the RSS, MSE, and RMSE would be calculated the same way as it is for the multiple linear regression model.

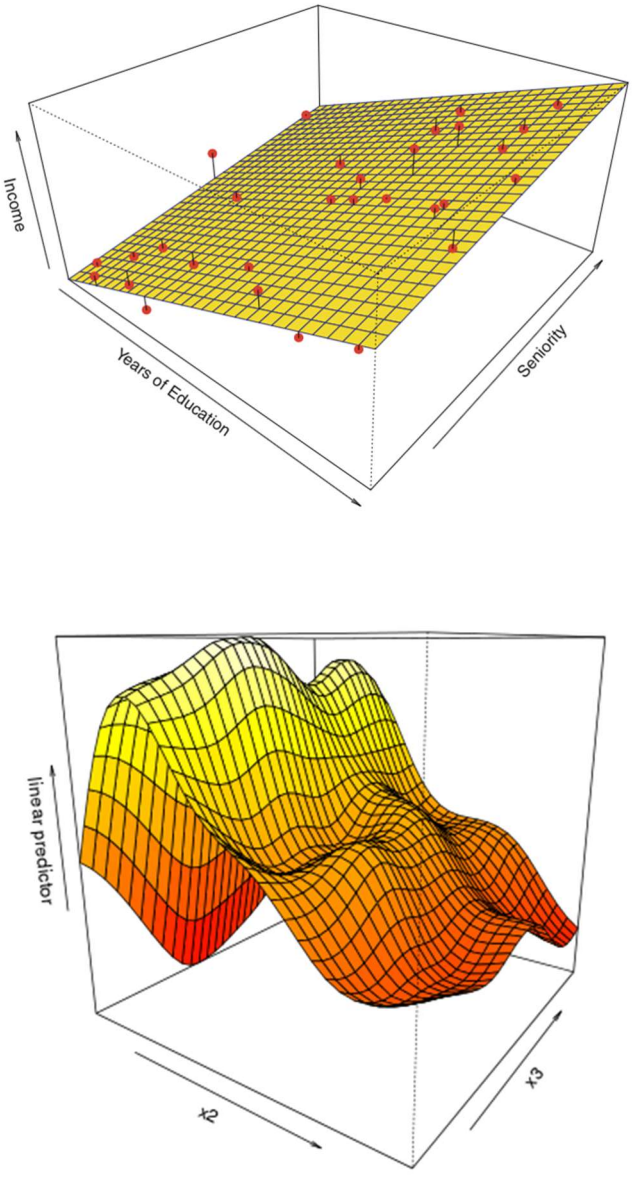

Figure 1: Multiple linear regression model consisting of two predictor variables (Years of Education and Seniority) adapted from James et al. (2017) and GAM consisting of two predictor variables (x2 and x3) fitted using spline functions (adapted from Logan, 2014).

The basis functions of a GAM, or those components in the regression equation that are additive in nature and used to account for the variability of the response variable, may take on different

forms. In some models, it may even make sense for some of the basis functions to be linear. This really depends on the character of the dataset and what functions can best be used to model the variability in the system (James et al., 2017).

When using spline basis functions, an important model tuning variable is the smoothing parameter (λ), i.e., how many degrees of freedom should be used to fit a model to the data? The greater the degrees of freedom, the "wigglier" a spline function will be. Cross-validation is a useful approach for optimizing the degrees of freedom in a way that minimizes the residual sum of squares (RSS) – remember that the average RSS is the MSE and the square root of the MSE is the RMSE. The details of cross-validation are described by James et al. (2017) – suffice to say cross-validation involves sampling the entire dataset to produce training datasets and fitting the function enough times to converge on a solution where RSS is minimized (James et al., 2017).

As with most parametric statistical learning techniques, validating the assumptions of the model is important. For GAMs, examining the residuals in detail using standard plots is the typical approach for conduction data validation. A Q-Q plot, predictors versus residuals, and a histogram of the residuals should be roughly linear (verifying normality), random, and normally distributed; respectively. A plot of fitted values versus the response variable should be approximately linear (Figure 2) (Laurinec, 2017).

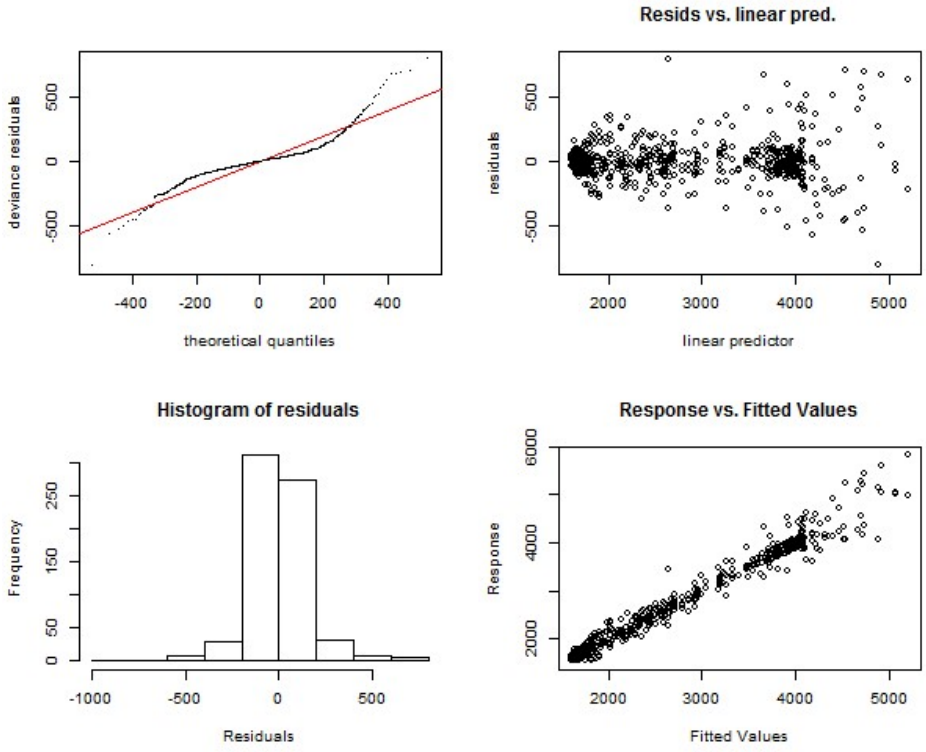

Figure 2: Model diagnostic plots for a GAM (adapted from Laurinec, 2017).

In this case, DEP considered an initial dataset of 148 contracts addressing 2,551 wells and executed between 1989 and 2018. All costs were inflation corrected using the Consumer Price Index. The open source statistical software R was used to complete the analysis (R Core Team, 2021).

First, the dataset was randomly divided into training and test datasets. Seventy-five percent of the cases were used to train the model and the remaining 25% of the cases were used to test the model. Several rounds of modeling and diagnostic work (Laurinec, 2017) were completed and cross-validation was used to optimize the degrees of freedom (i.e., tuning parameters) chosen for all spline basis functions used in the model (James et al., 2017).

It was found that per-well plugging costs ($\log_{10}$ transformed) could be reliably predicted as a function of the average depth of wells on a contract ($P<0.05$), the number of wells on a contract ($P<0.0001$), and the location (i.e., mean center longitude) of wells on a contract ($P<0.001$).

Pĩneiro et al. (2008) indicate that one common method for evaluating regression models of any type is to plot the predicted values (x-axis) versus the observed values (y-axis). Hess (2020) shows how uncertainty can be evaluated using such an approach. In Figure 3, predicted per-well costs versus observed per-well costs are shown for the test dataset used to evaluate the GAM's performance. The simple linear regression fit is depicted along with confidence and prediction intervals at a level of 95% and 75%, respectively. The relationship between the predicted per-well costs and observed per-well costs was found to be statistically significant ($P<0.000000001$) and the goodness-of-fit for the model is also reasonable ($R^2=0.6392$), supporting the effectiveness of the GAM for prediction purposes.

Examining Figure 3, it also is apparent that the GAM slightly underpredicts actual plugging costs, i.e., the linear model fit has a slope slightly steeper than unity (1:1). Additionally, use of the simple regression model allows for a thorough characterization of uncertainty, as associated confidence and prediction intervals can be reported for each prediction of future, observed per-well plugging costs. The 75% prediction interval also isolates all of the closely spaced test dataset points. Because of this, DEP has chosen to use output from the GAM along with the simple linear regression model and the 75% upper prediction limit to estimate plugging costs for the NOI.

Next, for the 17,367 wells classified as "Orphaned" in the prior section, the GAM was used to predict per-well plugging costs. For conservatism, all costs were estimated as single-well contracts and the longitude was used for the mean center longitude variable in the model.

Depth was not initially available for every well, and so a method was developed to estimate depth using publicly available data from the PA DCNR digital publication "Oil and gas fields and pools of Pennsylvania – 1859-2011" (Carter et al., 2015). Using ArcGIS Pro, wells were matched with the DCNR dataset by API number, and the recorded "Total Depth" was selected. For wells without individual documentation in the DCNR dataset, the depths were pulled from the "Average Production Depth" of either the overlapping pool or of the closest pool within a 1-mile radius of the well. For remaining wells (45) with no overlapping or nearby pools, the well depths were "hand-picked" by calculating the average depth for the 3 nearest wells and/or pools. Depending

on the geographic setting, “nearest” means radially or within the same surficial geologic formation.

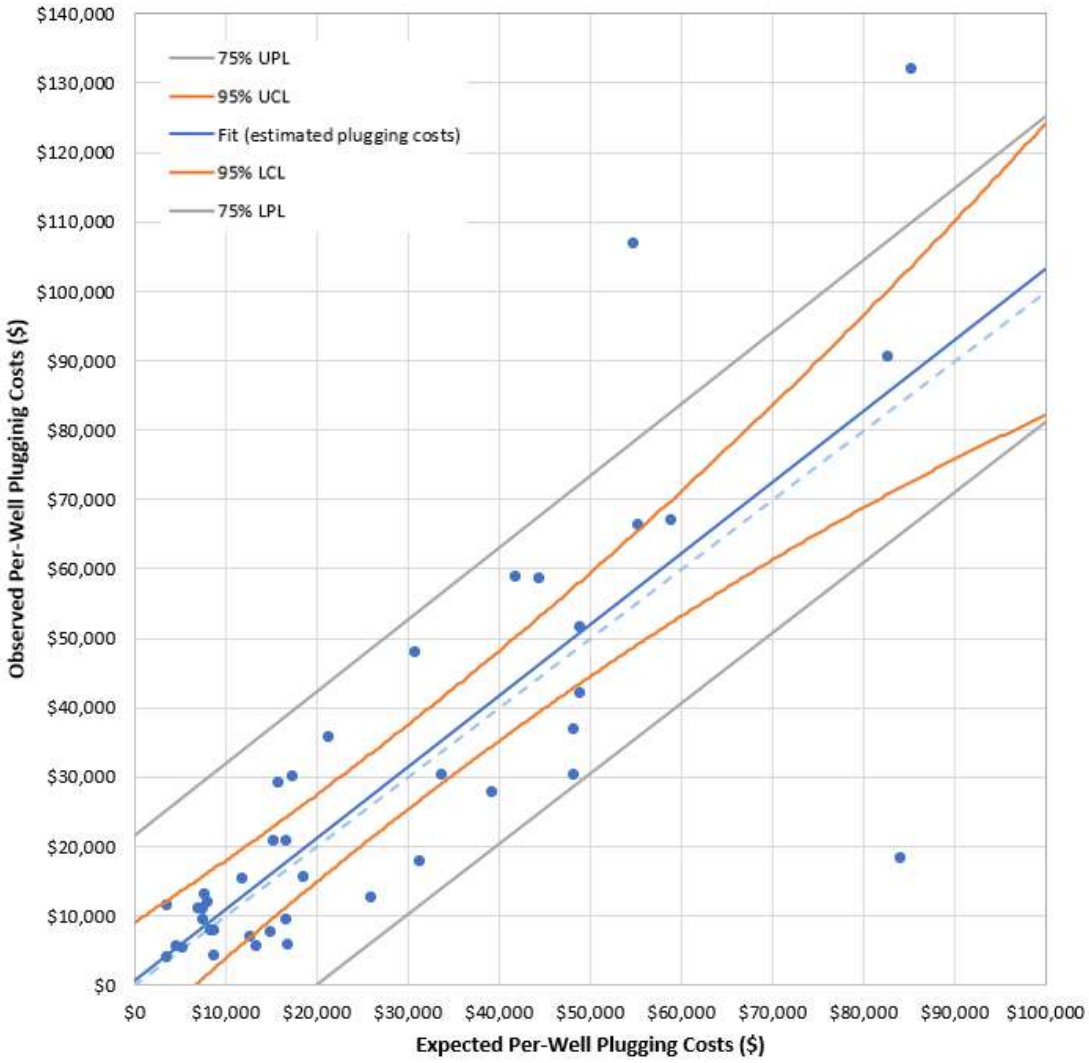

Figure 3: Expected versus observed plugging costs for GAM. 95% confidence and 75% prediction intervals are shown. The dashed light-blue line represents unity (1:1 slope).

For the 17,367 wells evaluated using the statistical models, the average 75% upper prediction is approximately $68,068 per well. For future contracts that have the characteristics used to build the model and given enough future contracts, 75 out of 100 times the observed per-well cost is not anticipated to be higher than this value. The derivation of this value is depicted in Figure 4.

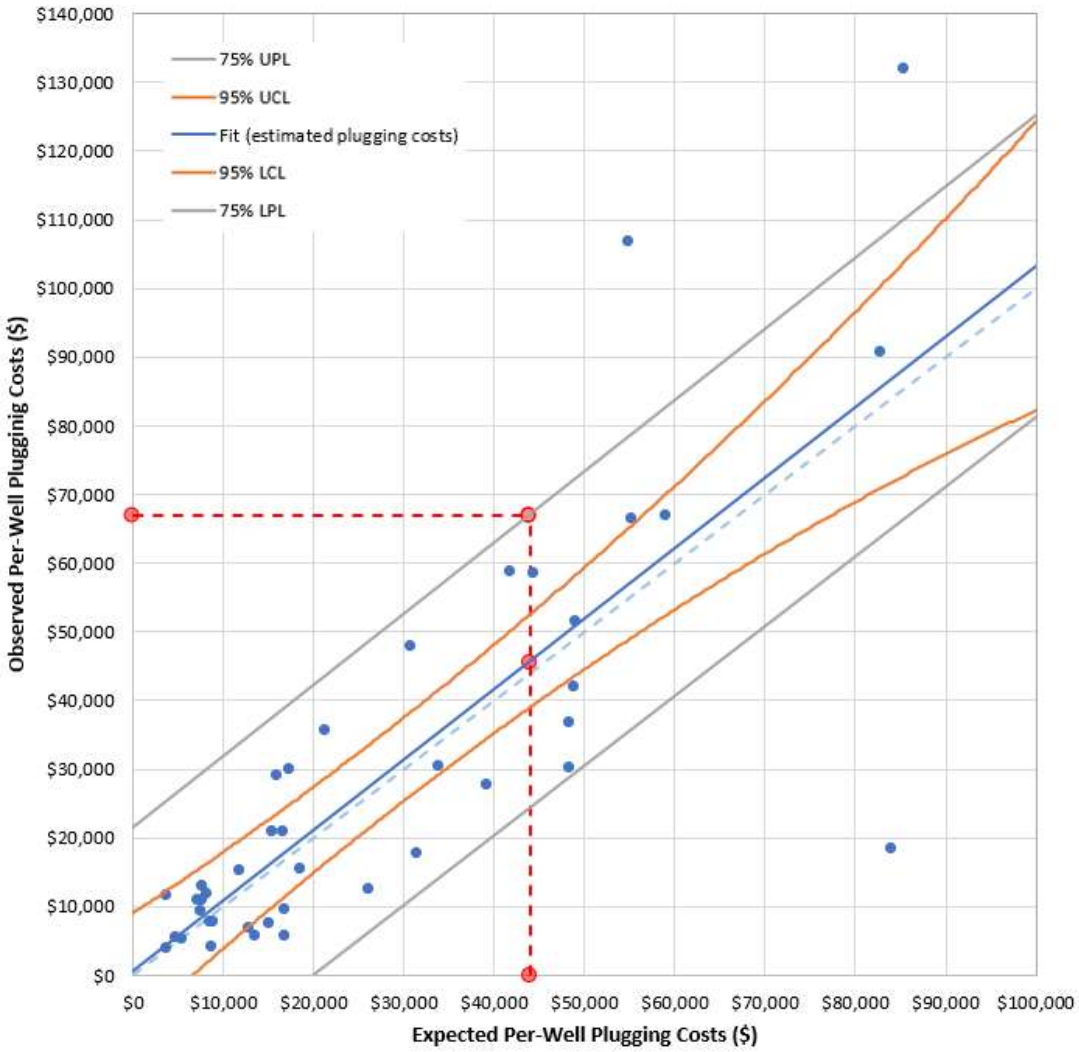

Figure 4: Expected versus observed plugging costs for GAM. 95% confidence and 75% prediction intervals are shown. The dashed light-blue line represents unity (1:1 slope). The red point intersecting the upper prediction limit function and extending to the y-axis represents the 75% upper prediction limit of the mean per-well plugging costs ($68,068). The red point along the x-axis is the mean expected per-well plugging cost derived from the GAM and the red point intersecting the blue "Fit" function is the mean observed per-well plugging cost. Note the figure is not to scale.

Figures 5 shows the model output for all 17,367 wells evaluated using the GAM and simple linear regression models, along with the 95% confidence and 75% prediction intervals for the simple linear regression model. The per-well costs have been ranked from lowest to highest.

Figure 6 is an accompanying distribution of the data shown in Figure 5.

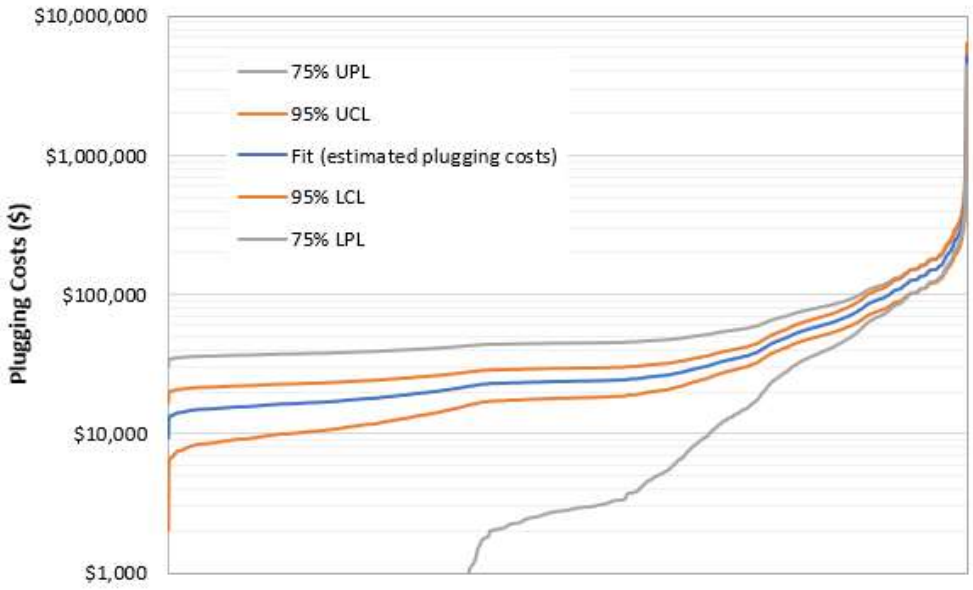

Figure 5: Model output for all 17,367 wells evaluated using the GAM and simple linear regression models, along with the 95% confidence and 75% prediction intervals for the simple linear regression model. The per-well costs have been ranked from lowest to highest.

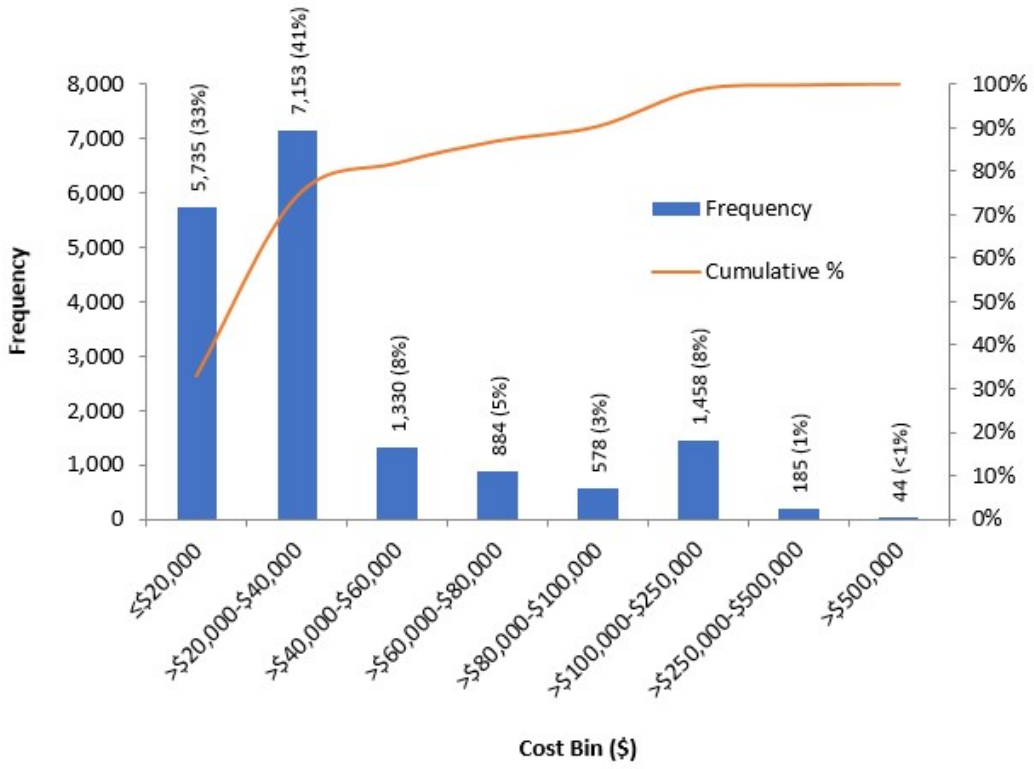

Figure 6: Distribution of model plugging costs for all 17,367 wells evaluated using the GAM and simple linear regression models.

The code used to develop all regression models is included in the appendix of this document. Access to raw data has also been provided.

As mentioned previously, 9,541 of the DEP reported "Orphaned" wells do not have associated locations. To estimate the projected costs to plug or reclaim these orphaned wells, reclaim adjacent land, and decommission and removed associated infrastructure, the 75% upper prediction limit mean per-well costs was scaled up using the following equation (Equation 1).

Equation 1

$$x = \$68{,}068\ per\ well * 26{,}908\ wells$$

where x = scaled-up costs for addressing 26,908 wells = $1,831,573,744

REFERENCES

1. James, G., Witten, D., Hastie, T., and Tibshirani, R., 2017, *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

2. Logan, M., 2014, Workshop 9.15.4: Non-linearity, URL http://www.flutterbys.com.au/stats/downloads/slides/pres.9.15.4.html#1.0, accessed December 7, 2021.

3. Laurinec, P., 2017, Doing magic and analyzing seasonal time series with GAM (Generalized Additive Model) in R, URL https://petolau.github.io/Analyzing-double-seasonal-time-series-with-GAM-in-R/, accessed December 7, 2021.

4. R Core Team, 2021, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

5. Pĩneiro, G., Perelman, S., Guerschman, J.P., and Paruelo, J., 2008, *How to evaluate models: Observed vs. predicted or predicted vs. observed?*, Ecological Modelling, pp. 316-322.

6. Hess, R., 2020, Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst, Nonlin. Processes Geophys, Vol. 27, Issue 4, URL https://npg.copernicus.org/articles/27/473/2020/.

7. Carter, K. M., Moore, M. E., Harper, J. A., and others, 2015, Oil and gas fields and pools of Pennsylvania—1859–2011: Pennsylvania Geological Survey, 4th ser., Open-File Report OFOG 15–01.2, 10 p., personal geodatabase and shapefiles.

APPENDIX: R CODE FOR PER-WELL PLUGGING COST PREDICTIONS

```
# NOTE THAT BOLD UNDERLINE TEXT INDICATES FILE PATHS THAT NEED TO BE ESTABLISHED BY USER DEPENDING ON WHERE THEY STORE SOURCE FILE INFORMATION

> library(splines)
> library(gam)
> library(car)
> library(caTools)
> AAA=file.path("FILE_PATH/Plugging_Costs.csv")
> AAA=read.csv(AAA,stringsAsFactors=FALSE)
> attach(AAA)
> names(AAA)
 [1] "Contract"                       "County"
 [3] "Municipality"                   "District"
 [5] "Group"                          "High_Cost"
 [7] "Inflation_Corrected_Cost_Per_Well"   "Inflation_Corrected_Contract_Amount"
 [9] "Contract_Days_Per_Well"         "Avg_Well_Depth"
[11] "Contract_Issue_Date"            "Last_Plug_Date"
[13] "Days"                           "Well_Count_Oil"
[15] "Well_Count_Gas"                 "Perc_Oil"
[17] "Other_Wells"                    "Total_Wells"
[19] "Contract_Year"                  "Mean_Cent_Lat"
[21] "Mean_Cent_Lon"                  "NW"
> set.seed(2)
> split=sample.split(AAA,SplitRatio=3/4)
> training_setAAA=subset(AAA,split==TRUE)
> test_setAAA=subset(AAA,split==FALSE)
> fitAAA=smooth.spline(Avg_Well_Depth,cv=TRUE)
> fitAAA$df
[1] 2.388601
> fit2AAA=smooth.spline(Total_Wells,cv=TRUE)
> fit2AAA$df
[1] 2.86306
> fit3AAA=smooth.spline(Mean_Cent_Lon,cv=TRUE)
> fit3AAA$df
[1] 6.204111
> GAMAAA=gam(Inflation_Corrected_Cost_Per_Well~s(Avg_Well_Depth,2.4)+s(Total_Wells,2.9)+
s(Mean_Cent_Lon,6.2))
> par(mfrow=c(3,1))
> plot(GAMAAA,se=TRUE,col="cadetblue")
> summary(GAMAAA)

Call: gam(formula = Inflation_Corrected_Cost_Per_Well ~ s(Avg_Well_Depth,
    2.4) + s(Total_Wells, 2.9) + s(Mean_Cent_Lon, 6.2))
Deviance Residuals:
   Min     1Q Median     3Q    Max
-54072 -13231  -2099   7764 191889
```

(Dispersion Parameter for gaussian family taken to be 796816351)

Null Deviance: 290524496403 on 147 degrees of freedom
Residual Deviance: 107968716786 on 135.5001 degrees of freedom
AIC: 3467.374

Number of Local Scoring Iterations: NA

Anova for Parametric Effects

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
s(Avg_Well_Depth, 2.4)	1.0	8.8402e+10	110.9439	< 2.2e-16	***
s(Total_Wells, 2.9)	1.0	5.7994e+09	7.2783	0.007867	**
s(Mean_Cent_Lon, 6.2)	1.0	3.1612e+09	3.9673	0.048403	*
Residuals	135.5	1.0797e+11	7.9682e+08		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects

	Npar Df	Npar F	Pr(F)	
(Intercept)				
s(Avg_Well_Depth, 2.4)	1.4	3.4589	0.050496	.
s(Total_Wells, 2.9)	1.9	5.0013	0.009034	**
s(Mean_Cent_Lon, 6.2)	5.2	8.6707	2.521e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
> GAM2AAA=gam(Inflation_Corrected_Cost_Per_Well~Avg_Well_Depth+s(Total_Wells,2.9)+
s(Mean_Cent_Lon,6.2))
> predAAA=predict(GAM2AAA)
> AAA$Pred=predAAA
> resAAA=residuals(GAM2AAA)
> AAA$Res=resAAA
> par(mfrow=c(1,1))
> qqPlot(AAA$Res)
[1] 106 139
> plot(AAA$Pred,AAA$Res)
> hist(AAA$Res,main="Histogram of Residuals",xlab="Residuals")
> plot(AAA$Pred,AAA$Inflation_Corrected_Cost_Per_Well)
> abline(0,1,col="red")
> AAA2=AAA[-c(106,139),]
> GAM3AAA=gam(log10(Inflation_Corrected_Cost_Per_Well)~s(Avg_Well_Depth,2.4)+
s(Total_Wells,2.9)+s(Mean_Cent_Lon,6.2),data=AAA2)
> summary(GAM3AAA)

Call: gam(formula = log10(Inflation_Corrected_Cost_Per_Well) ~ s(Avg_Well_Depth,
    2.4) + s(Total_Wells, 2.9) + s(Mean_Cent_Lon, 6.2), data = AAA2)
Deviance Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.627884 -0.158750  0.004374  0.157631  0.636260

(Dispersion Parameter for gaussian family taken to be 0.0578)

   Null Deviance: 31.2691 on 145 degrees of freedom
Residual Deviance: 7.7208 on 133.4998 degrees of freedom
AIC: 12.1367

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                    Df  Sum Sq Mean Sq  F value    Pr(>F)
s(Avg_Well_Depth, 2.4)   1.0 11.7996 11.7996 204.0252 < 2.2e-16 ***
s(Total_Wells, 2.9)      1.0  2.8634  2.8634  49.5101 9.326e-11 ***
s(Mean_Cent_Lon, 6.2)    1.0  0.1391  0.1391   2.4045    0.1234
Residuals            133.5  7.7208  0.0578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                  Npar Df  Npar F     Pr(F)
(Intercept)
s(Avg_Well_Depth, 2.4)     1.4  5.2058    0.0143 *
s(Total_Wells, 2.9)        1.9 10.3876 8.721e-05 ***
s(Mean_Cent_Lon, 6.2)      5.2  6.5749 1.265e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> pred2AAA=predict(GAM3AAA)
> AAA2$Pred2=pred2AAA
> res2AAA=residuals(GAM3AAA)
> AAA2$Res2=res2AAA
> par(mfrow=c(1,1))
> qqPlot(AAA2$Res2)
[1] 112 138
> plot(AAA2$Pred2,AAA2$Res2)
> hist(AAA2$Res2,main="Histogram of Residuals",xlab="Residuals")
> plot(AAA2$Pred2,log10(AAA2$Inflation_Corrected_Cost_Per_Well))
> abline(0,1,col="red")
> which(training_setAAA$ï..Contract=="BOGM-90-8")
[1] 76
> which(training_setAAA$ï..Contract=="BOGM-98-12")
[1] 101
> training_set2=training_setAAA[-c(76,101),]
> GAM4=gam(log10(Inflation_Corrected_Cost_Per_Well)~s(Avg_Well_Depth,2.4)+s(Total_Wells,2.9)+
s(Mean_Cent_Lon,6.2),data=training_set2)
> predGAM4=predict(GAM4,newdata=test_setAAA)
```

```
> test_setAAA$Pred3=predGAM4
> res3=log10(test_setAAA$Inflation_Corrected_Cost_Per_Well)-test_setAAA$Pred3
> test_setAAA$Res3=res3
> plot(test_setAAA$Pred3,log10(test_setAAA$Inflation_Corrected_Cost_Per_Well))
> abline(0,1,col="red")
> test_setAAA$PredTrans=10^(test_setAAA$Pred3)
> test_setAAA$ObsTrans=log10(test_setAAA$Inflation_Corrected_Cost_Per_Well)
> mean((test_setAAA$Pred3-test_setAAA$ObsTrans)^2)
[1] 0.05383905
> (0.05383905)^0.5
[1] 0.2320324
> PredTrainGAM= predict(GAM4,newdata=training_set2)
> training_set2$Pred=PredTrainGAM
> training_set2$PredTrans=10^(training_set2$Pred)
> training_set2$ObsTrans=log10(training_set2$Inflation_Corrected_Cost_Per_Well)
> mean((training_set2$Pred-training_set2$ObsTrans)^2)
[1] 0.05215996
> (0.05215996)^0.5
[1] 0.2283856
> summary(GAM4)
# all predictors are significant at 0.05 or less when fitted with spline basis functions – see Anova for
Nonparametric Effects
Call: gam(formula = log10(Inflation_Corrected_Cost_Per_Well) ~ s(Avg_Well_Depth,
    2.4) + s(Total_Wells, 2.9) + s(Mean_Cent_Lon, 6.2), data = training_set2)
Deviance Residuals:
      Min        1Q    Median        3Q       Max
-0.564189 -0.162292 -0.006683  0.144962  0.582104

(Dispersion Parameter for gaussian family taken to be 0.0592)

    Null Deviance: 24.3729 on 104 degrees of freedom
Residual Deviance: 5.4766 on 92.5 degrees of freedom
AIC: 14.8628

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                    Df Sum Sq Mean Sq  F value    Pr(>F)
s(Avg_Well_Depth, 2.4)  1.0 8.1272  8.1272 137.2670 < 2.2e-16 ***
s(Total_Wells, 2.9)     1.0 2.6221  2.6221  44.2866 1.975e-09 ***
s(Mean_Cent_Lon, 6.2)   1.0 0.1069  0.1069   1.8058    0.1823
Residuals            92.5 5.4766  0.0592
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Anova for Nonparametric Effects
                        Npar Df  Npar F     Pr(F)
(Intercept)
s(Avg_Well_Depth, 2.4)     1.4  3.7087 0.0431300 *
s(Total_Wells, 2.9)        1.9 12.3167 2.547e-05 ***
s(Mean_Cent_Lon, 6.2)      5.2  5.5554 0.0001312 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> write.csv(training_set2,file="Training_Data_GAM.csv")
> write.csv(test_setAAA,file="Test_Data_GAM.csv")
> attach(test_setAAA)
> linear_model=lm(Inflation_Corrected_Cost_Per_Well~PredTrans)
> summary(linear_model)
Call:
lm(formula = Inflation_Corrected_Cost_Per_Well ~ PredTrans)

Residuals:
   Min     1Q Median     3Q    Max
-68713  -8506    505   5705  49820

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  702.583   4352.198   0.161    0.873
PredTrans      1.026      0.121   8.477 2.21e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17720 on 39 degrees of freedom
Multiple R-squared:  0.6482,      Adjusted R-squared:  0.6392
F-statistic: 71.87 on 1 and 39 DF,  p-value: 2.21e-10

> CIs=predict(linear_model,interval="confidence",level=0.95)
> PIs=predict(linear_model,interval="prediction",level=0.95)
> write.csv(CIs,file="GAM_CI_linear_model.csv")
> write.csv(PIs,file="GAM_PI_linear_model.csv")
> PIs75=predict(linear_model,interval="prediction",level=0.75)
> write.csv(PIs75,file="GAM_PI_75_linear_model.csv")
# the next section of code imports new cost data
(Master_Sheet_AO_Depths_Stat_Ready_No_Fed_Loc.csv) in *.csv file, predicts costs per well using GAM
model, transforms predicted costs to $ and exports them to a new *.csv file
(PredictionsFormulaGrant.csv), predicts actual plugging costs using a simple linear regression model of
observed vs. expected values, estimates 95% confidence and prediction intervals and 75% prediction
interval, and exports all predicted values, confidence, and prediction intervals as *.csv files.
> BBB=file.path("FILE_PATH /Master_Sheet_AO_Depths_Stat_Ready_No_Fed_Loc.csv")
> BBB=read.csv(BBB,stringsAsFactors=FALSE)
> predGAMBBB=predict(GAM4,newdata=BBB)
> BBB$Pred2=10^(predGAMBBB)
```

```
> write.csv(BBB,file="PredictionsFormulaGrant.csv")
> FG=file.path("C:/Users/mipelepko/OneDrive - Commonwealth of
Pennsylvania/Desktop/Costs/PredictionsFormulaGrant.csv")
> FG=read.csv(FG,stringsAsFactors=FALSE)
> attach(FG)
> names(FG)
> PredFGCI=predict(linear_model,newdata = FG,interval="confidence")
> PredFGPI=predict(linear_model,newdata = FG,interval="prediction")
> write.csv(PredFGCI,file="CIPredictionsFG.csv")
> write.csv(PredFGPI,file="PIPredictionsFG.csv")
> PredFGPI75=predict(linear_model,newdata = FG,interval="prediction",level=0.75)
> write.csv(PredFGPI75,file="PIPredictions75FG.csv")
```