GISC 6387 Project: Predicting Oil Well Yields in Russia

By Rachel Brasier

.

Russia is the third-largest oil producing country in the world, after the United States and Saudi Arabia. Russian oil fields produce approximately 11.5 million barrels per day, or 11 percent of the world's total oil supply. A few circumstances make Russian oil production choices unique. First, with relatively low production costs, Russia can afford to maintain wells in lower-producing areas as well as high-producing areas. Second, much of Russia's natural resource extraction infrastructure was built under the Soviet Union, which often outweighed practicality with ideological and cultural concerns. The positioning of modern wells is at least somewhat dependent upon existing infrastructure. Third, Russia covers an expansive territory with greatly varying geological features and characteristics.

Therefore, the distribution of oil wells across Russia is a product of its current cost structure, its historical infrastructure, and its geology. This paper will attempt to predict the production of oil wells based on those factors, using production and location of wells with known oil production and geological features data from the United States Geological Survey (USGS).

I. Review of Literature

The 2000 World Petroleum Assessment conducted by the United States Geological Survey is considered one of the most comprehensive evaluations of worldwide oil and gas production capacity. One principle in the USGS methodology is that geological factors like tectonic history can affect petroleum reserves accumulated in a given area. Exploration history is also relevant; larger accumulations tend to have been found earlier than smaller accumulations; however, exploration trends also can be affected by economic, technological, and political factors (explored in Campbell 1968). Accumulation size is also skewed; there are far fewer large than small accumulations. USGS estimates are considered the standard for petroleum capacity estimates in other papers, such as Aguilera (2011).

Aguilera's paper uses a Variable Shape Distribution (VSD) model to calculate total petroleum endowment in petroleum provinces throughout the Former Soviet Union. The model uses known volumes of conventional petroleum reserves to calculate unknown volumes of conventional petroleum reserves, based on shape and size. The VSD model estimate for oil, natural gas, and NGL endowment in the assessed provinces very closely matches previous assessments by the United States Geological Survey (USGS) in their 2000 World Petroleum Assessment.

Hamida, et al. (2017) explore a variety of geometry-based optimization approaches for well placement in oil fields, including a measurement for quantitative

similarity between wells.

II. Study Area and Data Sources

This paper will combine data from a variety of sources. The United States Geological Survey (USGS) conducted an assessment of the geological provinces of the Former Soviet Union, which can be mapped to a shapefile of geological provinces and spatially joined with well location data.

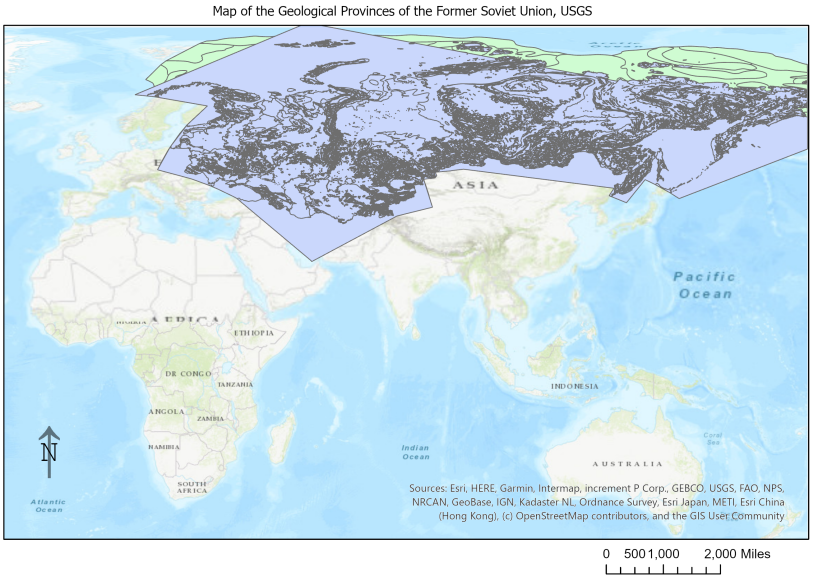

Figure 1. : Map of the geological provinces of the Former Soviet Union, USGS 1999.

The website energybase.ru provides the location of 558 currently operating oil and gas wells in Russia and Kazakhstan, which can be geocoded using coordinates extracted from the webpage HTML. Oil and gas production numbers are provided for 98 of these wells.

One loop captures the information printed in the data tables on the 28 web pages in the energybase.ru oil and gas field index, then appends them end-to-end in a single vector:

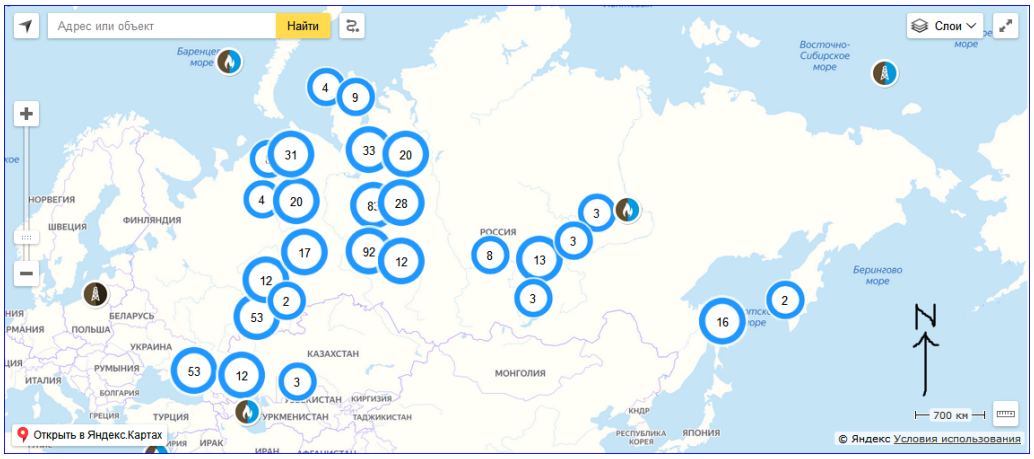

Figure 2. : Map of oil and gas wells in Russian Federation, energybase.ru (accessed June 22, 2020).

```
wellsdat <- NULL
for (pg in 1:28){
  webpage <- paste0("https://energybase.ru/oil-gas-field/index?page=",
                    pg) %>%
    read_html(encoding = "utf-8")
  out <- webpage %>%
    html_nodes("div.name, div.info") %>%
    html_nodes("a, small a, small, div.value") %>%
    html_text()
  out <- out[grepl("\n", out) == FALSE]
  wells <- as.data.frame.vector(out)
  wellsdat <- rbind(wellsdat, wells)
}
```

A set of nested if-statements then separates observations into "name," "company," "city," "field (deposit) type," and "production."

A separate loop stores the URL's for all 558 wells included in the data tables and appends them end-to-end in a single vector:

```
sites <- NULL
for (pg in 1:28){
  webpage <- paste0("https://energybase.ru/oil-gas-field/index?page=",
                    pg) %>%
    read_html(encoding = "utf-8")
  out <- webpage %>%
    html_nodes("div.name > a") %>%
    html_attr("href")
```

```
  site <- as.data.frame.vector(out)
  sites <- rbind(sites, site)
}
```

A final loop accesses the websites stored in the *sites* vector and stores the longitude and latitude coordinates from the embedded Yandex map:

```
<div class="item">
    <small>
        Координаты:
    </small>
    <br>
    <a href="#yandex-map">Широта: 45.000278</a><br><a href="#yandex-map">Долгота: 48.561944</a>          </div>
```

Figure 3. : Longitude and Latitude Coordinates from Yandex Map

```
for (i in 1:nrow(sites)){
  site <- sites$out[i]
  webpage <- paste0("https://energybase.ru", site) %>%
    read_html(encoding = "utf-8")
  out <- webpage %>%
    html_nodes("section.contacts div.item > a") %>%
    html_text()
  coord <- as.data.frame(cbind(out[1], out[2]))
  coords <- rbind(coords, coord)
}
colnames(coords) <- c("lat", "long")
coords$id <- as.numeric(row.names(coords))
coords$lat <- str_extract(coords$lat, "[[:digit:]].*$")
coords$lat <- as.numeric(coords$lat)
coords$long <- str_extract(coords$long, "[[:digit:]].*$")
coords$long <- as.numeric(coords$long)
```

III. Methodology

Because of the limited availability of production data for gas, gas-condensate, gas-oil, oil-gas, and oil-gas-condensate wells, this project will consider only oil wells. Production is known for 54 oil wells in the sample, and 195 of the geocoded oil wells have unknown production. This paper uses a variation on the classic inverse-distance weighting formula:

$$p_i = \frac{\sum_{j=1}^{n} \frac{1}{d_{ij}^2} p_j}{\sum_{j=1}^{n} \frac{1}{d_{ij}^2}} \tag{1}$$

where p_i is the predicted petroleum yield for well i, p_j is known petroleum yield for well j, and d_{ij} measures the distance between two wells i and j.

Table 1—: Russian classification of field and deposit types

Тип месторождения (залежи)	**Field type (deposits)**	**Translated Description**
Газовое (Г)	Gas	Only free gas.
Газоконденсатное (ГК)	Gas-condensate	Gas with condensates.
Газонефтяное (ГН)	Gas-oil	Oil and gas; the bulk of deposits are oil, and the gas cap does not exceed the volume of the oil part of the reservoir.
Нефтяное (НГ)	Oil	Oil only, saturated in varying degrees of gas.
Нефтегазовое (Н)	Oil-gas	Gas and oil; gas deposits with an oil rim and deposits in which the gas cap exceeds the volume of the oil part of the reservoir.
Нефтегазоконденсатное (НГК)	Oil-gas-condensate	Oil, gas, and condensates.

Source: Ministry of Natural Resources and Ecology of the Russian Federation, 2016.

I will estimate production in wells with known location but unknown production based on their similarity to wells with known oil production from energybase.ru.

The *rdist.earth* method in the *phangorn* package in *R* calculates the great-circle distance between each pair of wells i and j, assuming a radius of 6,366.71 km. A point-by-point comparison with the National Oceanic and Atmospheric Association's Longitude and Latitude Distance Calculator finds the same results over a sample of wells.

```
writeDist(rdist.earth(oil.mtrx, miles = FALSE, R = 6366.71),
     file = "oil.mtrx.csv")
oil.mtrx <- read.delim("oil.mtrx.csv", header = FALSE, sep = " ")
oil.mtrx <- as.matrix(oil.mtrx[2:250,2:250])
```

This process generates a $249x249$ symmetrical matrix A. To account for slight

Table 2—: Data Availability by Petroleum Type

Petroleum Type	Known Production	Unknown Production	Total Wells	Percent Known
Gas	0	49	49	0%
Gas-condensate	4	77	81	5%
Gas-oil	3	4	7	43%
Oil	54	195	249	22%
Oil-gas	12	48	60	20%
Oil-gas-condensate	22	82	104	21%

Table 3—: Description of Variables

Variable Name	Description
	Dependent Variable
Predicted oil production	Oil production predicted by model, million tonnes
	Independent Variables
Inverse distance	Inverse great-circle distance, kilometers
Geological type match	Indicator variable; 1 if wells i and j share geological province type; 0 otherwise
Known oil production	Oil production, if known, in million tonnes

errors in the calculation due to rounding, I replaced every cell on the diagonal with 0, to represent a distance of 0 km between well i and well i.

```
for(i in 1:dim(oil.mtrx)[1]) {oil.mtrx[i,i] = 0}
```

$$A_{249,249} = \begin{pmatrix} 0 & 1209.03 & 346.43 & 347.53 & \cdots & a_{1,249} \\ 1209.03 & 0 & 1149.13 & 861.74 & \cdots & a_{2,249} \\ 346.43 & 1149.13 & 0 & 398.66 & \cdots & a_{3,249} \\ 347.53 & 861.74 & 398.66 & 0 & \cdots & a_{4,249} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{249,1} & a_{249,2} & a_{249,3} & a_{249,4} & \cdots & a_{249,249} \end{pmatrix}$$

Then the inverse squared distance is calculated as $D = \frac{1}{A \cdot A}$, leaving zeroes where $A_{i,j}$ equals zero:

```
mtrx.dist.inv <- ifelse(oil.mtrx!=0, 1/(oil.mtrx*oil.mtrx), oil.mtrx)
dist.inv <- as.data.frame(mtrx.dist.inv)
```

$$D_{249,249} = \begin{pmatrix} 0 & 6.84e-7 & 8.33e-6 & 8.27e-6 & \cdots & d_{1,249} \\ 6.84e-7 & 0 & 7.57e-7 & 1.35e-6 & \cdots & d_{2,249} \\ 8.33e-6 & 7.57e-7 & 0 & 6.29e-6 & \cdots & d_{3,249} \\ 8.27e-6 & 1.35e-6 & 6.29e-6 & 0 & \cdots & d_{4,249} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{249,1} & d_{249,2} & d_{249,3} & d_{249,4} & \cdots & d_{249,249} \end{pmatrix}$$

Table 4—: Count of Oil Wells by Geological Province Type

Geology Type (Abbr.)	Geology Type	Known Production	Unknown Production	Total Wells
Pg	Paleogene	98	33	131
P	Permian	5	37	42
K	Cretaceous	2	35	37
N	Neogene	3	6	9
SEA	Sea	2	6	8
QT	Quaternary & Tertiary	4	3	7
J	Jurassic	1	5	6
Tr	Triassic	2	2	4
D	Devonian	1	1	2
O	Ordovician	1	0	1
H20	Water	0	1	1
Cm	Cambrian	0	1	1

Matrix G, also $249x249$ in dimensions, and symmetrical, notes if wells i and j share the same geological province type. Cells $G_{i,j}$ are coded 1 if wells i and j have the same geological province type, e.g. "Permian," and 0 if they have different types.

```
mtrx.glg <- matrix(NA, nrow = 249, ncol = 249)
for (i in 1:nrow(oil.wells)){
  for (j in 1:nrow(oil.wells)){
    mtrx.glg[i,j] <- ifelse(
      oil.wells$GLG[i] == oil.wells$GLG[j], ifelse(
        i != j, 1, 0),
     0)
  }
```

```
}
```

$$G_{249,249} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & g_{1,249} \\ 0 & 0 & 0 & 0 & \cdots & g_{2,249} \\ 0 & 0 & 0 & 1 & \cdots & g_{3,249} \\ 0 & 0 & 1 & 0 & \cdots & g_{4,249} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{249,1} & g_{249,2} & g_{249,3} & g)249,4 & \cdots & g_{249,249} \end{pmatrix}$$

A $249x1$ matrix O is created from the *prod* vector stored in the *oil.wells* dataframe. Missing values are replaced with zeroes.

```
prod <- as.matrix(oil.wells$prod)
prod[is.na(prod)] <- 0
```

$$O_{249,1} = \begin{pmatrix} 360 \\ 350 \\ 232 \\ 220 \\ \vdots \\ o_{249,1} \end{pmatrix}$$

A production weights matrix V recodes O so that missing values are replaced with zeroes, and any nonzero production numbers are recoded as 1. This way, cells in D and G corresponding to the zeroes in matrix O will not be counted in the final weights vector W.

```
prod.weights <- prod
prod.weights[is.na(prod.weights)] <- 0
prod.weights[prod.weights > 0] <- 1
```

$$V_{249,1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ v_{249,1} \end{pmatrix}$$

The $249x1$ weights matrix W is the product of matrices D, G, and V:

$$W_{DGV} = DGV; W_{DV} = DV; W_{GV} = GV \tag{2}$$

$$W_{DGV} = DGV = \begin{pmatrix} 0.06 \\ 0.01 \\ 0.14 \\ 2.35 \\ \vdots \\ w_{249,1} \end{pmatrix}; W_{DV} = DV = \begin{pmatrix} 7.18e-4 \\ 9.71e-5 \\ 8.28e-4 \\ 1.28e-3 \\ \vdots \\ w_{249,1} \end{pmatrix}; W_{GV} = GV = \begin{pmatrix} 3 \\ 0 \\ 32 \\ 32 \\ \vdots \\ w_{249,1} \end{pmatrix}$$

Meanwhile, predicted production for the oil wells is captured in three variants of $249x1$ matrix P, where P_{DGO} is a product of matrices D, G, and O; P_{DO} is a product of only matrices D and O, leaving out the geological province match matrix; and P_{GO} is a product of only matrices G and O, leaving out the inverse distance matrix.

$$P_{DGO} = \frac{DGO}{W_{DGO}}; P_{DO} = \frac{DO}{W_{DO}}; P_{GO} = \frac{GO}{W_{GO}} \tag{3}$$

DGO, DO, and GO are all $249x1$ vectors, so they can be divided respectively by W_{DGO}, W_{DO}, and W_{GO}, also $249x1$ vectors.

$$P_{DGO} = \begin{pmatrix} 61.74 \\ 108.48 \\ 61.08 \\ 60.89 \\ \vdots \\ p_{249,1} \end{pmatrix}; P_{DO} = \begin{pmatrix} 31.56 \\ 63.91 \\ 40.15 \\ 86.95 \\ \vdots \\ p_{249,1} \end{pmatrix}; P_{GO} = \begin{pmatrix} 21.67 \\ NA \\ 55.53 \\ 55.93 \\ \vdots \\ p_{249,1} \end{pmatrix}$$

```
dist.glg.weighted.pred <- mtrx.dist.inv %*% mtrx.glg %*% prod
dist.glg.pred <- dist.glg.weighted.pred/dist.glg.weights.all
```

IV. Results

I compare P, predicted production of each well, to O, the known production. This comparison includes the summary statistics of each formula's predicted oil production, as well as the mean squared error (MSE) for wells with known oil production – i.e. how well the models perform on the test set. There are more missing observations for the P_{GO} method based solely on geological province match because as the above table "Count of Oil Wells by Geological Province Type" shows, three of the wells do not have a geological match in the data set. The MSE is smallest for the P_{DGO} model, which only includes wells with the same geological province type, weighted by distance. However, the mean is the furthest from the actual mean, and the minimum is quite high and the maximum quite low (108.48 compared to 360.00). The P_{GO} model performs surprisingly well, implying

Table 5—: Comparing Predicted to Known Oil Production

	P_{DGO}	P_{DO}	P_{GO}	O
Minimum	8.33	4.17	4.27	1.80
Mean	57.01	64.78	61.15	63.87
Median	60.83	60.98	60.91	30.50
Maximum	108.48	334.53	350.00	360.00
Non-missing obs.	249	249	244	54
Missing obs.	0	0	5	195
Total obs.	249	249	249	249
MSE	6057.38	6533.70	6320.78	–
Error N	54	54	51	–

that the geological province type does play quite a large role in determining oil production.

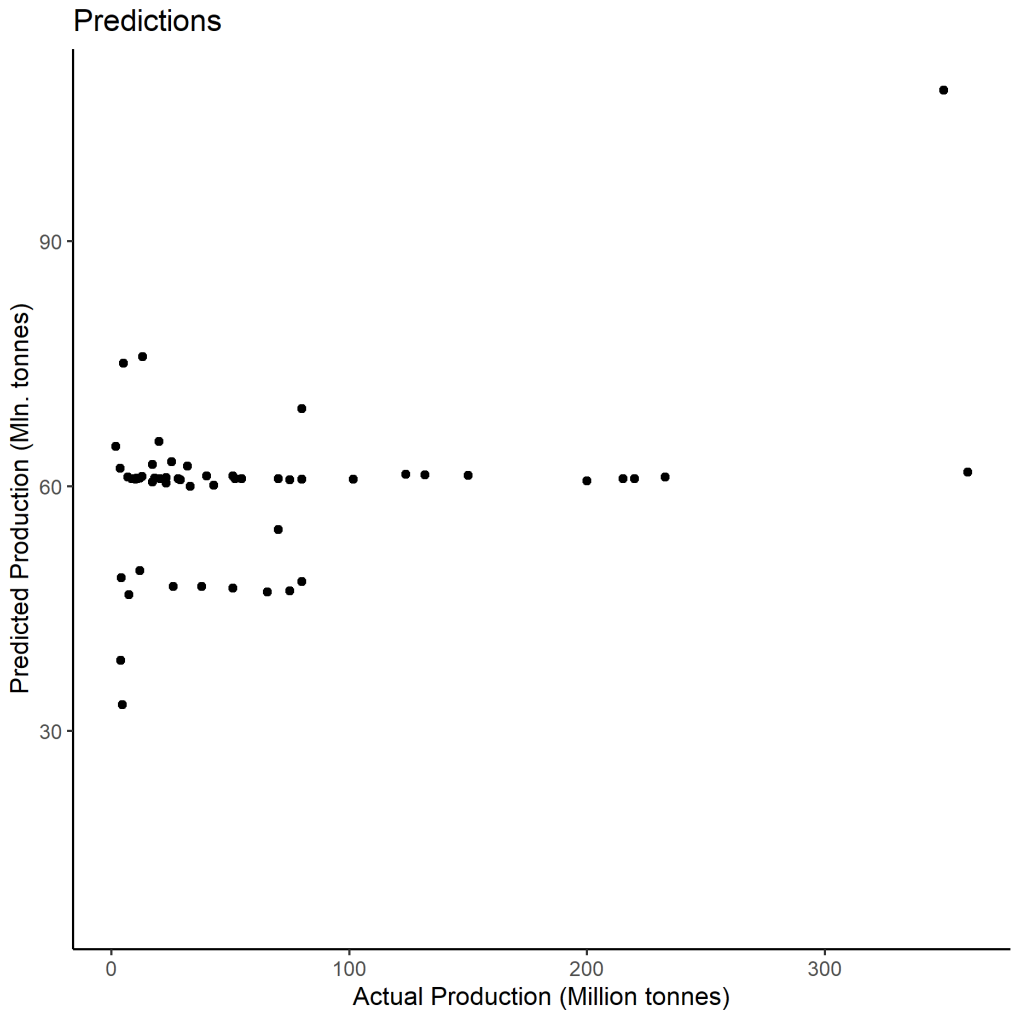

Figure 4. : Predicted oil prediction by well using P_{DGO} model, for 54 oil wells with known production.

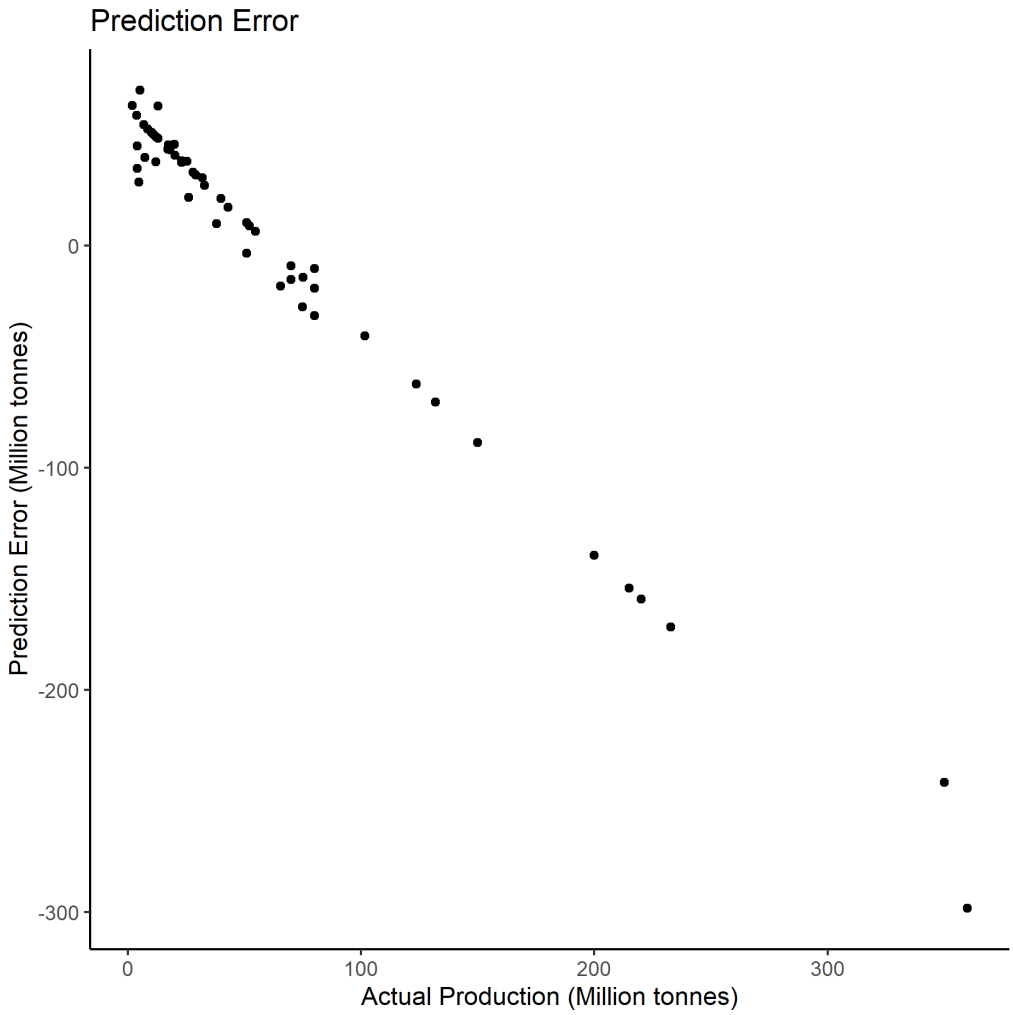

Figure 5. : Error in predicted oil production by well using the P_{DGO} model, for 54 oil wells with known production.

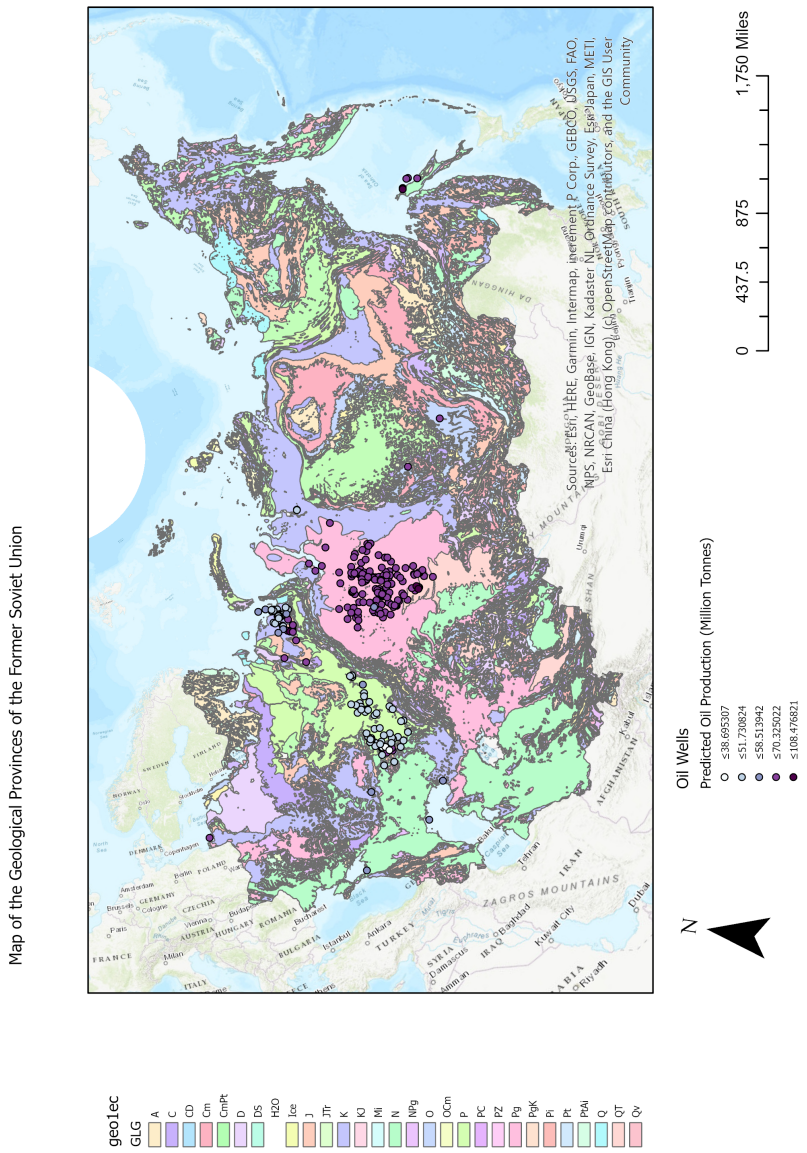

Figure 6. : Predicted oil production from P_{DGO}.

V. Conclusion

The accuracy of the models is lacking. The model weighting production estimates by distance and geological province type, P_{DGO}, is the best-performing in terms of mean squared error, but has a higher minimum value and much lower maximum value. The errors are particularly pronounced for high-producing wells, probably because of the inherent skew of the well production data.

Some directions that might be helpful for future research are training the model by feeding output predicted production back into the model as "known" production. The model would then have a full set of 249 wells to include in the weighted average.

Furthermore, some application of the VSD model like in Aguilera (2011) could be used to integrate data from the USGS production capacity estimates, which might help to improve accuracy of the model.

References

Aguilera, R. F. (2011). Modeling petroleum resources in provinces of the Former Soviet Union. Energy Exploration and Exploitation, 29(4), 379-396.

Campbell, Robert. Economic Reform in the USSR. The American Economic Review, May, 1968, Vol. 58, No. 2, Papers and Proceedings of the Eightieth Annual Meeting of the American Economic Association (May, 1968), pp. 547-558.

Charpentier, R. R., and Klett, T. R. (2005). Guiding principles of USGS methodology for assessment of undiscovered conventional oil and gas resources. Natural Resources Research, 14(3), 175-186.

Hamida, et al. (2017). An efficient geometry-based optimization approach for well placement in oil fields. Journal of Petroleum Science and Engineering, 149, 383-392.

Sources

Catalog of oil and gas wells in Russia (Каталог нефтяных и газовых месторождений России). Energybase.ru. Retrieved June 22, 2020, from https://energybase.ru/oil-gas-field.

Persits, F.M., Ulmishek, G.F., and Steinshouer, D.W., 1999, Maps showing geology, oil and gas fields and geologic provinces of the former Soviet Union: U.S. Geological Survey Open-File Report 97-470-E, 13 p., https://doi.org/10.3133/ofr97470E.