Assessing the Local Developmental Impact of Hydrocarbon Exploitation in a Mature Region: A Random Forest Approach

Vlad-Cosmin BULAI
The Bucharest University of Economic Studies, Romania
vldbulai@yahoo.com

Alexandra HOROBEȚ
The Bucharest University of Economic Studies, Romania
alexandra.horobet@rei.ase.ro

Abstract

The impact of natural resource exploitation has been a controversial topic, subject to intense debate. The literature has traditionally focused on its consequences on national socioeconomic development. More recently, scholars concentrated on local effects following greater availability of data at the subnational and project level. We add to the literature by concentrating on Romanian oil and gas operations, a mature region with a long history of hydrocarbon activities. Such regions have seldom been studied and we argue that in light of the ongoing energy transition these should garner greater interest, particularly those located within the European Union where environmental pressure is significant. Our methodology consists of testing the ability of the random forest classification algorithm to distinguish between local communities with oil and gas operations present and those without on a number of indicators which could be broadly considered developmental. The algorithm fails to accurately classify hydrocarbon-intensive communities, indicating that there are no significant differences between these and the rest. We argue that this is likely due to the limited tax collection powers of local governments, with royalties going directly to the central government with no specific distribution provision at the local level. Another potential explanation may be the diversification of local economies and existing related manufacturing and services activities.

Keywords: random forest; resource curse; hydrocarbon; extractive industry; local development;

JEL Codes: O13; Q32; Q35;

DOI: http://doi.org/10.24818/ejis.2019.05

1. Introduction

The impact of natural resource exploitation has been a topic of much debate and controversy, even without taking into account environmental and social dimensions. Economic benefits such as fiscal revenues and job creation have been weighed against the volatility of these incomes mirroring commodity cycles, the crowding out of other sectors, and the potential for conflict and the deterioration of institutions caused by rent-seeking behaviour. This is part of the so-called "resource curse" literature. More recently, scholars have begun to look at the effects at the local community level. This is also the aim of our paper. We focus on Romania due to data availability and because mature regions have seldom enjoyed academic attention. Furthermore, Romania presents an interesting characteristic in this regard since the oil and gas industry is highly geographically dispersed, but not present in all regions. This allows for a thorough assessment of the differences between resource and non-resource exploiting communities.

Our research is motivated by the European Union’s ambition to become climate-neutral by 2050, known as the Green Deal (European Commission, 2019). Populations in hydrocarbon-intensive communities are expected to be the most affected, in particular those in the less developed Eastern part. This fact is recognised by European policymakers with a Just Transition Mechanism established to support the transition in the most energy intensive and fossil fuel reliant regions. Romania is allocated around 10% of the total 7.5 billion EUR funding. This is the third largest share after Poland and Germany (European Commission, 2020) and illustrates the magnitude of the problem for Romania and the rest of Europe. A thorough assessment of the characteristics of the most affected communities should be of paramount importance to policymakers and the academic community, and a matter of some urgency. The likelihood that many fossil fuel reserves may become stranded due to environmental concerns is not a new idea. It has been the cornerstone of the so-called “carbon bubble” hypothesis (Ritchie and Dowlatabadi, 2015). In the present paper, we are concerned with oil and natural gas, but it would be reasonable to expect that coal would be among the first and most affected. Low quality coal such as lignite used in power generation is particularly vulnerable given mounting pressure to phase out coal from the power sector. Future research should pursue this area.

The assessment is carried out using economic and social indicators pertaining to all local administrative units (LAUs) in Romania. We present these below. The level of oil and gas activity is measured by the number of oil and gas wells present in a given LAU. This is our dependent variable. However, using a regression to explore the relationship between it and the indicators would make little sense, as we do not expect a linear relationship to be present. In other words, there is no reason to expect any differences between a LAU with one well and one with two wells. We thus turn our problem into a classification problem and apply the random forest algorithm. The rationale is that if there are differences between the LAUs with hydrocarbon extraction activities present and the ones where such activities are absent, in terms of the indicators chosen, the model should be able to distinguish between them. If it cannot do so accurately, then we can conclude that the impact is negligible or non-existent.

2. Literature Review

As previously noted, the resource curse literature has a long history and has traditionally been concerned with the country-level effects of natural resource wealth through the “Dutch disease” phenomenon. This represents the crowding out of other sectors, such as manufacturing, in the wake of a boom by depriving these of resources (resource movement effect) and through inflation generated by the surge in income (spending effect) (Corden and Neary, 1982). The theory has been backed by empirical evidence showing that countries with abundant natural resources tend to register lower economic growth (Sachs and Warner, 1997). This has received criticism, for example from Brunnschweiler and Bulte (2008) who make the distinction between resource dependence and abundance and find that abundance is positively correlated with growth. Despite criticism, the “Dutch disease” phenomenon has been studied at the regional level and found to occur in Canada’s mineral-rich regions (Papyrakis and Raveh, 2014). This highlights the importance of subnational-level analysis since even if the countrywide outcome is positive, the extractive industry may be an enclave within its region and generate only limited benefits for the local population. As an example, Aroca (2001) shows that the Chilean mining sector in one of the country's most prolific regions is largely disconnected from that region's economy.

The relatively recent availability of highly granular data relating to local government spending and transfers from the central authority, as well as various socioeconomic indicators and variables pertaining to individual resource extraction projects has allowed scholars to shift their focus to the community level (Cust and Poelhekke, 2015). The effects analysed go beyond the economic dimension of employment and income to include social outcomes such as crime. This review of the literature is not exhaustive but meant to give the reader an overview of the most relevant issues and findings which relate to the current study.

There has been significant attention given to the local impact of a natural resource boom in the case of both developed and developing countries. Ellem and Tonts (2018) chart the local impact of the global commodities boom and its aftermath on Australia, the Pilbara region in particular. The gains from the ramp up of mining activities were found to be limited by the transient nature of the workforce. Perry and Rowe (2015) also point to the negative impacts of a large transient workforce on local communities due to pressure on housing prices and local services, as well as the potential for crowding out other sectors such as tourism. Many of the same problems are identified for the Bowen Basin's coal extraction operations, with benefits tending to accrue more in larger urban centres while small communities support many of the costs (Rolfe *et al.*, 2007). In terms of local government finances, Drew, Dollery and Blackwell (2018) use a panel regression model to explore the relation between local administration expenditure and the number of mining companies in New South Wales, Australia. They find a mismatch between the taxes received and the costs imposed by mining operations on local governments.

The US unconventional hydrocarbon (shale) boom has also received significant consideration. Its contribution to incomes in Pennsylvania is found to come primarily from mineral rights ownership through leases and royalties, with wages playing a modest part (Hardy and Kelsey, 2015). Results are confirmed by Schafft *et al.* (2018) who further note the negative effects on the impoverished population from the spike in housing prices generated by the boom. The vulnerability of resource-rich local communities to external shocks had been documented even before last decade's shale boom. Oil market conditions were found to be significant determinants of employment levels in two Louisiana parishes which relied extensively on offshore oil extraction during the 1970s and 1980s (Gramling and Freudenburg, 1990).

In other resource rich areas, such as Peru, the presence of mineral extraction operations is found to benefit local income and poverty reduction, but the distribution of benefits is uneven (Loayza and Rigolini, 2016). In Brazil, municipalities with oil wealth tend to have worse outcomes in terms of development and higher inequality compared to similar ones lacking resources (Lima-de-Oliveira and Alonso, 2017).

3. Data and Methodology

3.1. The data

Our objective is to see whether LAUs with hydrocarbon exploitation activities and those without can be distinguished based on a set of 15 variables for year 2017. These are presented in Table 1.

Table 1. Development indicators

Indicator	Code	Unit	Source	Notes	Relation to development
Household natural gas consumption	HGC	Cubic meters per capita	National Institute of Statistics		**Positive** *(a high level would indicate availability of personal boilers)*
Welfare expenditures	WEL	RON per capita	Ministry of Finance	From local government budget	**Negative** *(a high level would indicate more persons in need of welfare)*
Civil servant salaries	CSAL	RON per capita	Ministry of Finance	The amount is per entire population. As such, a high number may represent an excess of government clerks	**Negative** *(this is somewhat tenuous, but a high level may indicate an inefficient local public administration with potentially negative consequences on development)*
Vehicle tax	VT	RON per capita	Ministry of Finance	From natural persons to local government	**Positive** *(a high level may point to either greater vehicle ownership or ownership of vehicles of greater value)*
Building tax	BT	RON per capita	Ministry of Finance	From natural persons to local government	**Positive** *(same as above)*
National testing graduation rate	NT	Percentage share	Ministry of Education	This is the test following completion of secondary schooling. A high rate may indicate a better living environment. We use this instead of high school graduation rates since education up to this point is mandatory.	**Positive** *(better results may indicate better living conditions)*
Persons in correctional facilities	JAIL	Persons per 1000 capita	National Institute of Statistics	By home address, not location of correctional facility	**Negative** *(more incarcerated persons would correspond to more crime)*
Educational expenditure	EDU	RON per pre-university school population	National Institute of Statistics	From local government budget	**Positive** *(a high level means funding is available)*
Infant mortality rate	MORT	Per 1000 live births	National Institute of Statistics		**Negative** *(may be correlated with the state of local healthcare facilities)*
Household water consumption	HWC	Cubic meters per capita	National Institute of Statistics		**Positive** *(access to water is a prerequisite for modern living conditions)*
Number of medical doctors	DOC	Per 1000 capita	National Institute of Statistics		**Positive** *(access to healthcare is a vital necessity)*
Living space	SPACE	Square meters per capita	National Institute of Statistics		**Positive** *(a high level may indicate better living conditions)*
Population density	DENS	Inhabitants per square kilometre	National Zoning and Real Estate Publicity Agency (ANCPI) for area and Ministry of Regional Development and Public Administration for population	Population divided by surface area	**Positive** *(high density corresponds to a high degree of urbanization)*
Number of PCs per classroom	PC	Number	National Institute of Statistics		**Positive** *(a high number would indicate that funding is available and that schools are well-equipped)*
Average number of employees	EMP	Per capita	National Institute of Statistics		**Positive** *(high levels of employment point to a well-developed local economy)*

Source: Authors

The indicators can be broadly interpreted as developmental. In many cases, this relationship is self-explanatory (e.g. high welfare expenditures would indicate a greater level of poverty). We note our interpretation of the developmental nature of each variable and admit that for some this is debatable. The rationale for data selection has primarily been availability. Indicators traditionally used in the literature, such as wages and income inequality, are not available at this level in Romania. We partially mitigate this by using taxes on assets (buildings and vehicles) collected by local governments as a proxy for wealth. The analysis is exploratory in nature since, based on the literature, no prior hypothesis can be formed regarding the relationship between the level of activity and each indicator or any general concept of development. In other words, LAUs with hydrocarbon operations might not enjoy greater employment or livings standards since the employment may go to the large cities and the local government's tax collection powers are limited. Rent-seeking behaviour may lead to an inefficient local public administration with a high civil servant expenditure relative to the population. On the other hand, this behaviour may be limited or non-existent as the local government's powers are limited and royalties are paid to the central budget with no amounts diverted to the communities where resources are being exploited.

The level of activity is determined based on the number of wells drilled between 1970 and 2018. The data come from Wood Mackenzie. We could go farther back in time, but we wish to avoid selecting LAUs with long-abandoned operations. In any case, onshore resources are still mainly bound to the historical regions. In order to avoid having the codification of the extent of operations appear arbitrary we use four different classifications (Table 2).

Table 2. Different classifications

Classification	Levels of activity – number of wells
4L	No =0; low (0, 10]; average (10, 20]; high >20
6L	No =0; very low (0, 5]; low (5, 10]; average (10, 15]; high (15, 20]; very high >20
2L	No =0; yes >0
2L5	No [0, 5]; yes >5

Source: Authors

There are 3181 LAUs in the country, of which 2860 are communes, the rest are towns and municipalities. In order to mitigate the issue of heterogeneity, as towns and cities may be more developed and benefit from more diversified local economies, we apply the analysis on both the restricted sample of communes and on the full sample. Communes are typically rural, but some heterogeneity is expected to be present with some located around large cities potentially enjoying better living standards. The same classifications are used for both cases. These are informed by the histograms in Figures 1.a) and 1.b).

Figure 1.a) Wells distribution for all LAUs

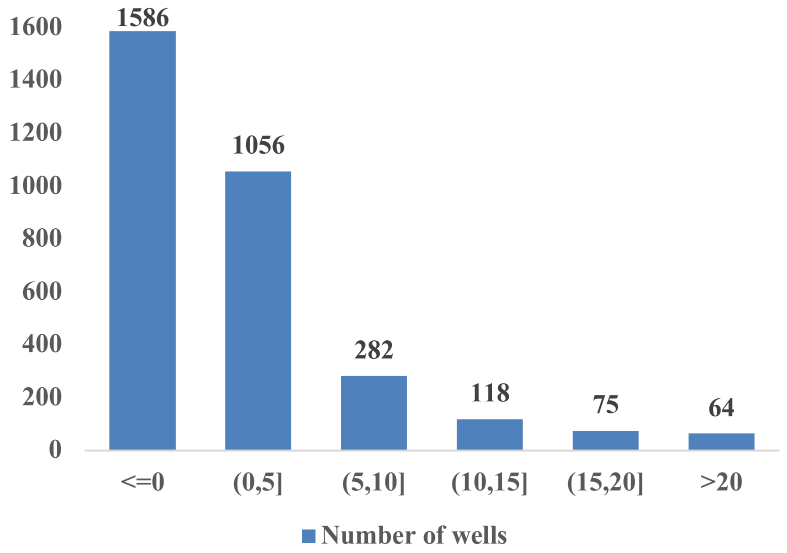
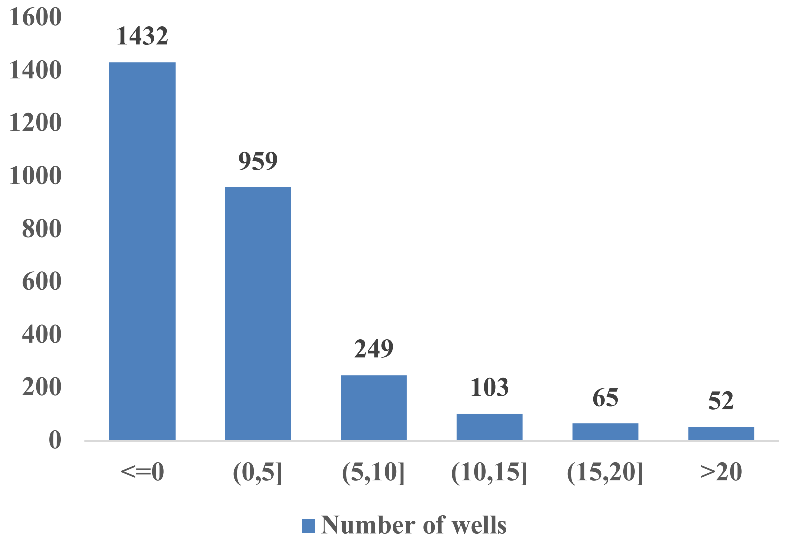


Figure 1.b) Wells distribution for communes



Source: Authors' calculations, based on Wood Mackenzie data

3.2. The random forest algorithm

There are a number of classification algorithms such as support vector machines and boosted tree ensembles. Computations were done in Matlab which allows for testing a number of these with default parameters and selecting the most accurate. The random forest (bagged decision tree ensemble) proved the most accurate, but in many cases the gains were small.

The random forest algorithm was developed by Breiman (2001) and consists of an ensemble of decision trees which, in the case of a classification problem, vote to determine membership to a given class. The solution is therefore the most popular class for each item. The trees are constructed by bootstrap aggregation of samples (bagging) and random feature selection. Bagging was introduced by Breiman (1996) as a way to improve accuracy by repeatedly sampling observations with replacement and using these to construct a number of classifiers (trees), the results of which are then aggregated. Random feature selection means that the split at each node of each tree is based on the best variable among a random subset of all variables considered.

An overview of the methodology for the construction of decision trees, also known as the CART (Classification and Regression Trees) algorithm, is given by Steinberg (2009). Trees are formed by starting at a root node containing all observations and partitioning the data at each branch according to a splitting rule. The rule is determined by a goodness-of-split metric such as Gini impurity. This is the probability of incorrectly classifying a randomly chosen observation if it were randomly labelled according only to the class distribution of the entire data. The objective is thus to minimize impurity, the minimum value being zero and indicating a perfectly homogenous group (all values belonging to a single class). The Gini impurity is evaluated for each branch and a weighted sum (with the number of observations) is computed and subtracted from the parent node Gini impurity. This is the Gini gain. A higher value indicates lower impurity remaining. Therefore, the best split is the one which produces the highest gain. The process is repeated resulting in a structure resembling an upside-down tree.

The accuracy of the random forest model is determined by keeping part of the data outside the bootstrap sample (out-of-bag) and running the tree on the out-of-bag data. On average around a third of observations are kept out-of-bag. The predictions made by each tree are aggregated and used to compute the out-of-bag error rate.

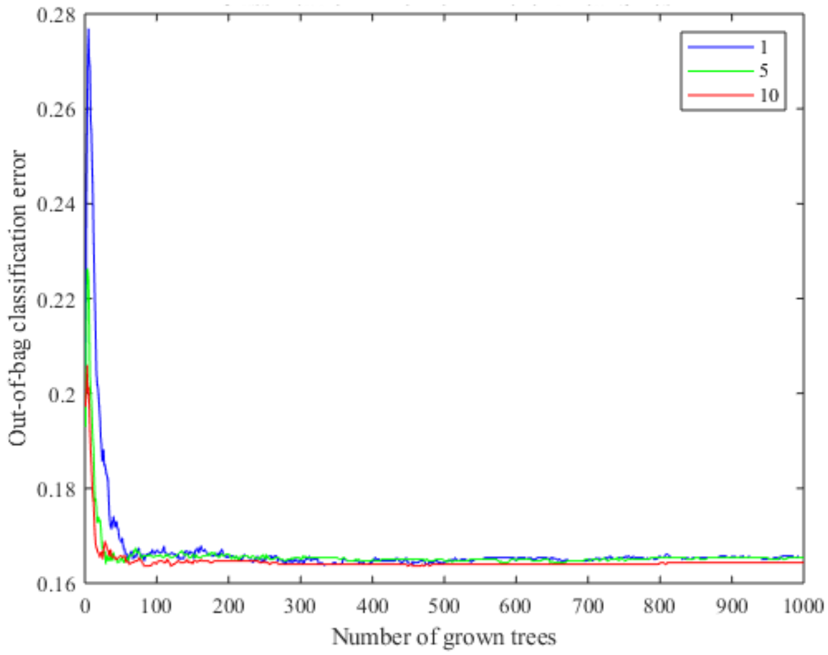
The advantage of ensemble methods over single decision trees lies in the greater accuracy, but at the cost of interpretability (Loh, 2014). The random forest algorithm is also user-friendly in that it is not very sensitive to its parameters – the number of trees and the number of variables used for the random feature selection (Liaw and Wiener, 2002).

4. Results and Discussion

The results for all four activity level classifications (Table 2) are similar in both the full and restricted samples. We therefore present only the last case (2L5) for the restricted sample. Having only two classes simplifies the illustration, and labelling LAUs with more than five wells as hydrocarbon exploiting allows us to test the classification strength for LAUs located closer to the centre of the hydrocarbon extraction regions.

We first pick the number of trees and minimum leaf size (number of observations in each terminal node) based on out-of-bag classification error. Running the algorithm for up to 1000 trees with three different minimum leaf sizes (Figure 2) yields a minimum error (0.1636) at 82 trees and a minimum leaf size of 10. These parameters are used going forward, but the minimum leaf size does not appear to be particularly important and the error rate stabilizes after less than 100 trees. Nevertheless, less trees means faster execution so there is no reason to use a larger number than necessary.

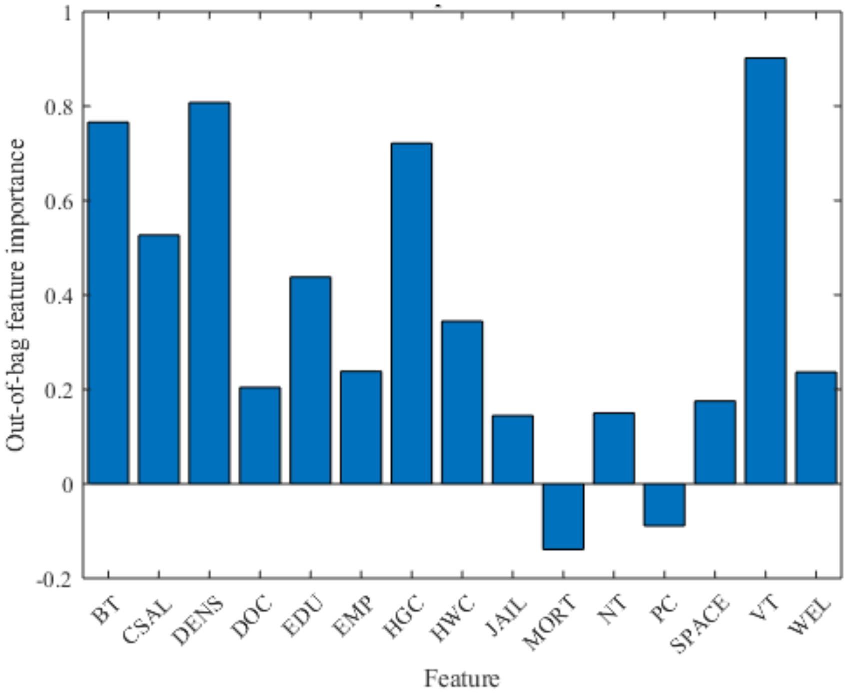
Figure 2. Classification error for different leaf sizes



Source: Authors' calculations

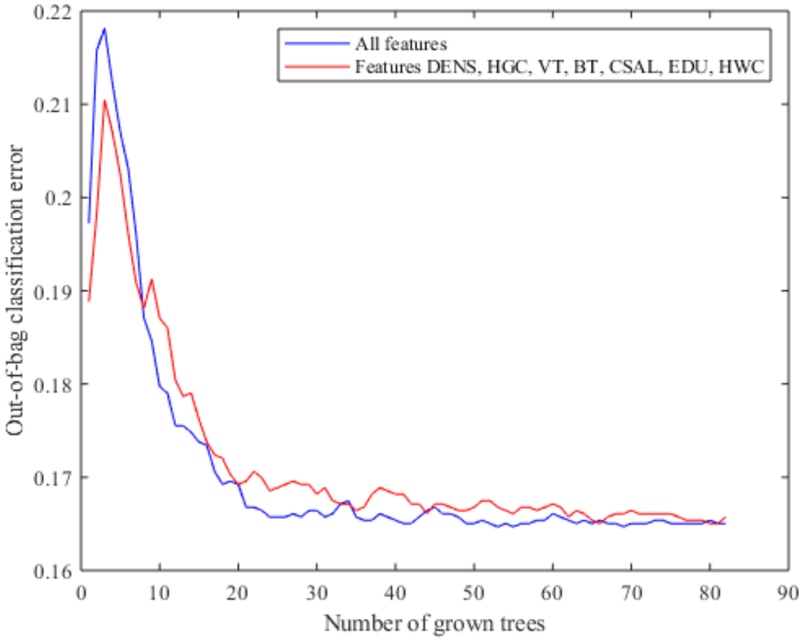
Next, we evaluate the importance of each variable (Figure 3). This represents the out-of-bag error increase when its values are randomly permuted. A high value means that the variable is influential in determining the classification result. Only a few of the variables stand out so we check the accuracy of the model using only the seven most important ones against the full set. In Figure 4, it can be seen that for our forest of 83 trees the performance is virtually identical. This restricted set of variables is used going forward.

Figure 3. Variable importance



Source: Authors' calculations

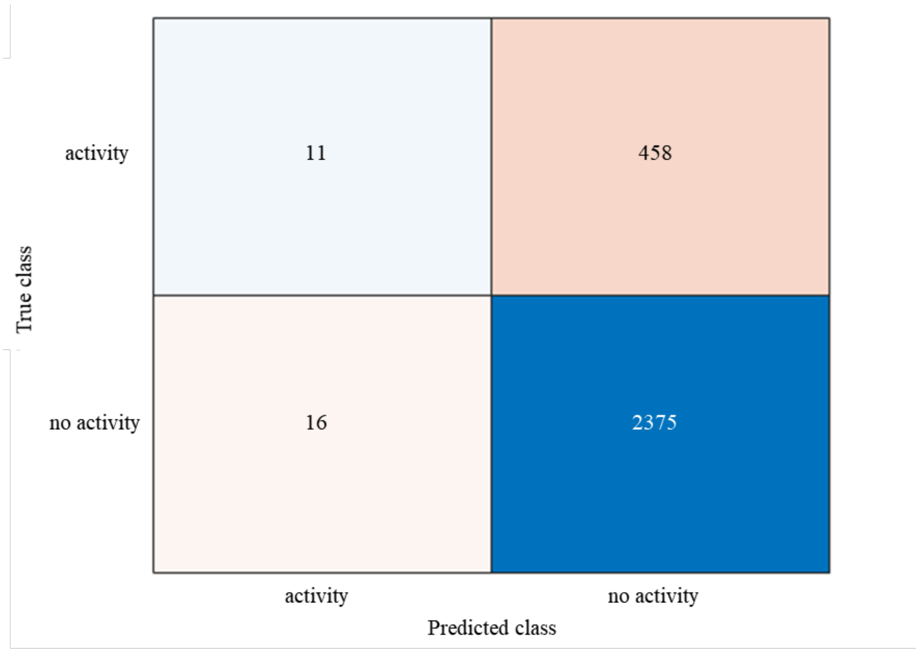
Figure 4. Classification error for different sets of variables



Source: Authors' calculations

While the vehicle and building tax, population density and household gas consumption have been identified as among the most important, the actual performance of the model is yet to be determined. The relatively low error rate or ~16% would indicate that it is accurate, but this metric may be misleading since the class distribution is unbalanced. There are 2391 LAUs with no activity and only 469 with activity. The confusion matrix (Figure 5) shows that the model labels the former with good accuracy, but fails at classifying the latter. If we consider the "activity" case as positive and the "no activity" case as negative, our model has a high specificity of 99% (true negative predictions divided by total negative cases), but a recall of only 2% (true positive predictions divided by total positive cases).

Figure 5. Confusion matrix



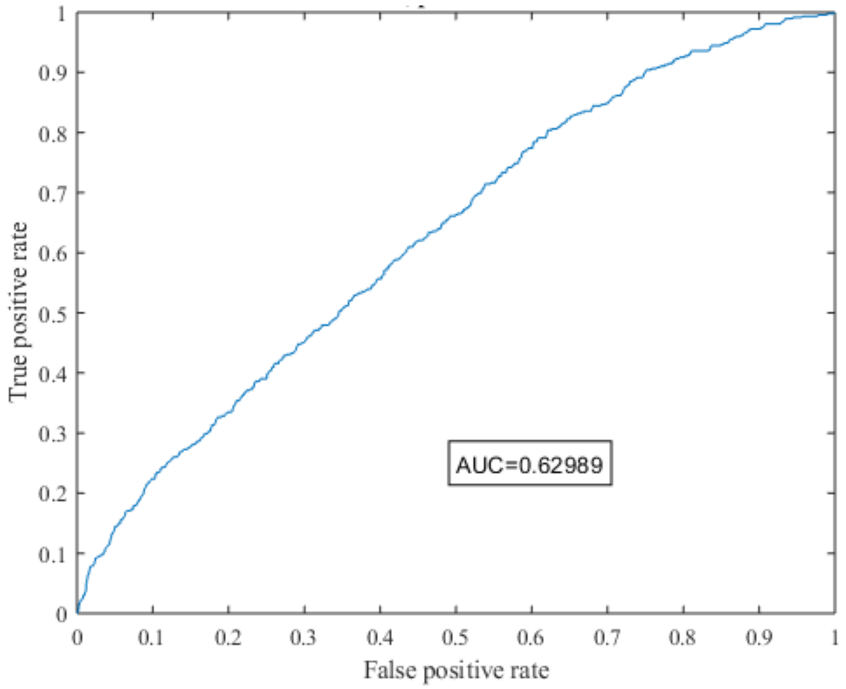
True class	activity	no activity
activity	11	458
no activity	16	2375

Predicted class

Source: Authors' calculations

The receiver operating characteristic (ROC) curve is another useful illustration of the model's performance. This plots the true positive rate against the false positive rate for each observation. The accuracy can be determined by computing the area under the curve (AUC). The perfect model would have and AUC of one, indicating that its classifications are 100% accurate. The worst model would have an AUC of 0.5, represented as a diagonal line. This translates to the model's performance being no better than a random classification – a coin toss. As we can see in Figure 6, our model's AUC is only 0.63, with a curve close to the diagonal.

Figure 6. ROC curve



Source: Authors' calculations

Given the poor performance of the model, we can conclude that there are no significant differences between LAUs with hydrocarbon operations and the rest, in terms of the indicators available. As previously noted, the results are similar for the full sample and across all classifications. However, the analysis has a major limitation given by its cross-sectional nature as the data used is only for year 2017.

5. Conclusions

The goal of this paper has been to assess the local developmental impact of the oil and gas industry in Romania. This is a mature hydrocarbon region with a long history. Furthermore, connected sectors are also present in the country, such as refineries and equipment manufacturing. Our assessment is based solely on identifying differences between LAUs with hydrocarbon operations present and those without. Because of this and the fact that we do not take offshore operations into account, our analysis likely understates the local importance of the industry.

The analysis has been exploratory in nature. The 15 indicators used can be described as developmental, but their selection was primarily based on data availability. The random forest model failed to correctly classify the LAUs by hydrocarbon activity level using the indicators. The conclusion that can be drawn is that there are no significant differences between local communities in hydrocarbon intensive regions and the rest. However, the analysis has a major limitation due to its cross-sectional nature. Despite its limitations, this study has presented evidence toward the neutral local developmental impact of the hydrocarbon industry in Romania. We posit that this is due in part to the taxation system. Most taxes, including resource royalties, are collected by the central government. There is no provision to distribute part of the royalties to local governments. Other potential determinants include the integration with downstream activities (refining, electricity generation) and the presence of related services and manufacturing activities.

To our knowledge, this is the only study of this type carried out for a mature region with an extremely long history of hydrocarbon extraction. The focus on a mature region adds to a literature which is mostly concerned with the effects of a boom following the discovery of large resource deposits. This is undoubtedly of interest, as are the long-term effects. It is possible that in a sufficiently diversified economy, the long-term local developmental impact will tend to be neutral. We hesitate to make generalisations since there are too many idiosyncrasies given by the long history of the industry in the country. Future research should concentrate on expanding the analysis to other mature regions and other types of resources.

Finally, it should be noted that if the developmental impact is negligible, the overall impact may be negative as local communities are exposed to environmental risks. While we believe that these should receive some form of compensation, we do not argue for a mechanism to distribute royalties to local government budgets. This is because the quality of Romania’s public administration, at a local level in particular, may be insufficient to ensure that the extra funds are allocated properly instead of disappearing under mysterious circumstances. This is not just a theoretical possibility. Brazil has such a mechanism which transfers royalties to municipalities situated on the coast since oil is extracted offshore. This has been found to increase revenues and spending considerably, but the quality improvements to public services have been modest (Caselli and Michaels, 2013). We therefore suggest that funding should be allocated toward specific local projects in partnership with local authorities and the oil companies operating in these areas. European funding should be accessed especially in the area of energy transition.

References:

Aroca, P. (2001). Impacts and development in local economies based on mining: the case of the Chilean II region. *Resources Policy*, *27*(2), 119-134.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brunnschweiler, C. N., Bulte, E. H. (2008). The resource curse revisited and revised: A tale of paradoxes and red herrings. *Journal of environmental economics and management*, *55*(3), 248-264.

Caselli, F., Michaels, G. (2013). Do oil windfalls improve living standards? Evidence from Brazil. *American Economic Journal: Applied Economics*, *5*(1), 208-38.

Corden, W. M., Neary, J. P. (1982). Booming sector and de-industrialisation in a small open economy. *The economic journal*, *92*(368), 825-848.

Cust, J., Poelhekke, S. (2015). The local economic impacts of natural resource extraction. *Annu. Rev. Resour. Econ.*, *7*(1), 251-268.

Drew, J., Dollery, B. E., Blackwell, B. D. (2018). A square deal? Mining costs, mining royalties and local government in New South Wales, Australia. *Resources Policy*, *55*, 113-122.

Ellem, B., Tonts, M. (2018). The global commodities boom and the reshaping of regional economies: the Australian experience. *Australian Geographer*, *49*(3), 383-395.

European Commission. (2019). *The European Green Deal*. Brussels. COM/2019/640 final

European Commission. (2020). *JTM and JTF Allocation Table*. Brussels. Retrieved February 5, 2020, from https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_66

Gramling, R., Freudenburg, W. R. (1990). A Closer Look at "Local Control": Communities, Commodities, and the Collapse of the Coast 1. *Rural Sociology*, *55*(4), 541-558.

Hardy, K., Kelsey, T. W. (2015). Local income related to Marcellus shale activity in Pennsylvania. *Community Development*, *46*(4), 329-340.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Lima-de-Oliveira, R., Alonso, M. L. (2017). Fueling development? Assessing the impact of oil and soybean wealth on municipalities in Brazil. *The Extractive Industries and Society*, *4*(3), 576-585.

Loayza, N., Rigolini, J. (2016). The local impact of mining on poverty and inequality: evidence from the commodity boom in Peru. *World development*, *84*, 219-234.

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329-348.

Papyrakis, E., Raveh, O. (2014). An empirical analysis of a regional Dutch Disease: the case of Canada. *Environmental and Resource Economics*, *58*(2), 179-198.

Perry, M., Rowe, J. E. (2015). Fly-in, fly-out, drive-in, drive-out: The Australian mining boom and its impacts on the local economy. *Local Economy*, *30*(1), 139-148.

Ritchie, J., Dowlatabadi, H. (2015). Divest from the carbon bubble? Reviewing the implications and limitations of fossil fuel divestment for institutional investors. *Review of Economics and Finance*, *5*(2), 59-80.

Rolfe, J., Miles, B., Lockie, S., Ivanova, G. (2007). Lessons from the social and economic impacts of the mining boom in the Bowen Basin 2004-2006. *Australasian Journal of Regional Studies, The*, *13*(2), 134.

Sachs, J. D., Warner, A. M. (1997). Fundamental sources of long-run growth. *The American economic review*, *87*(2), 184-188.

Schafft, K. A., McHenry-Sorber, E., Hall, D., Burfoot-Rochford, I. (2018). Busted amidst the Boom: The Creation of New Insecurities and Inequalities within Pennsylvania's Shale Gas Boomtowns. *Rural Sociology*, *83*(3), 503-531.

Steinberg, D. (2009). CART: classification and regression trees. In *The top ten algorithms in data mining* (pp. 193-216). Chapman and Hall/CRC.