Predicting Petroleum Fields in Ethnic Regions with Social and Economic Data: Evidence from Africa (Poster)

Kweku Opoku-Agyemang*
Algorithmic Fairness and Opacity Group, UC Berkeley and Development Economics X
kao75@cornell.edu

ABSTRACT

The paper develops an artificial neural network that predicts the presence of petroleum fields within ethnic country regions across sub-Saharan Africa using rich socioeconomic microdata. Using data from around 300,000 households from 1997 to 2014, the model accurately predicts the presence of petroleum fields in ethnic regions with an overall accuracy of 89.7%. Furthermore, the accuracy of the test and validation were found to be 89.9%. The slightly-increased accuracy in predicting petroleum fields suggests that socioeconomic data may be complementary to standard petroleum studies approaches in unpacking the social context of oil. The paper also explores dimensionality reductions to optimally characterize, organize, and visualize the data. Social science data may have a helpful role to play for oil resources and sustainable development

CCS CONCEPTS

• **Applied computing**; • **Machine learning**; • **Social and professional topics**;

KEYWORDS

oil, sustainable development, Africa

ACM Reference Format:
Kweku Opoku-Agyemang. 2021. Predicting Petroleum Fields in Ethnic Regions with Social and Economic Data: Evidence from Africa (Poster). In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '21), June 28–July 02, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3460112.3471971

1 INTRODUCTION

This paper asks: can sociodemographic and economic data sufficiently predict the presence of petroleum fields in ethnic regions? Much oil discoveries are based on petroleum science, and social science may or may not eventually play a complementary role in predicting the presence of petroleum fields within particular ethnic regions. This paper uncovers some correlations of socioeconomic variables with oil resources. This approach is important for two reasons. First, much work in development economics focuses on the negative impacts of natural resources [1], and second, such results are also seen in developed countries [2]. Both citizenry and industry alike have roles to play for sustainable development prior to oil discoveries [3].

The paper applies an artificial neural network that predicts the presence of petroleum fields within ethnic-country localized regions. Using data from almost 300,000 households that did or did not live within oil regions across Africa from 1997 to 2014, a neural network is able to accurately predict the presence of petroleum fields in ethnic regions with an overall accuracy of about 89.7%. Further, the accuracy of the test and validation were found to be around 89.9%. The slightly-increased accuracy in the detection of petroleum fields suggests that socioeconomic data may be complementary in unpacking the social context of natural resources. To provide context for the findings, the paper also applies a newer dimensionality reduction technique to optimally characterize, organize, and visualize social science data given its highly non-linear structure, noise, and continuous progressive nature in panel datasets.

In part because deep learning models are based on neurobiological analogies, I also attempt to use a machine learning-based nonlinear dimensionality reduction approach commonly used in biology to visualize the data. These are a class of statistical methods that visualize high-dimensional data by giving each datapoint a location in a two or three-dimensional map (the paper focuses on two dimensions). In the deep learning literature, these are often done in one of two ways. First, a t-distributed stochastic neighbor embedding or tSNE [4], preserves local structure, but its disadvantage is that the global structure is traded off and therefore not preserved. An approach meant to improve on t-SNE is the Uniform Manifold Approximation and Projection or UMAP approach [5]. It is similar to tSNE, but preserves more of the underlying global structure of data than the t-SNE does. However, UMAP improves on global, but not local data structure.

The method used in this paper, documented to improves on these two approaches, is the Potential of Heat diffusion for Affinity-based Transition Embedding (PHATE) approach [6]. The PHATE is a visualization method that captures both local and global nonlinear structure using an information-geometric distance between data points. Relative to the tSNE and UMAP approaches, PHATE consistently preserves a range of patterns in data, including continual progressions, branches and clusters [6]. I apply it to the sociodemographic dataset here in visualization exercises, building further on the biological analogy that supports deep learning models.

The paper taps into a growing literature that uses algorithms to find correlations and make predictions that add business value and context for oil and gas pipeline field operations. Machine learning has long been used to target resources in the oil industry, as have conventional production methods in Venezuela [7]. Experienced

*Many thanks to the COMPASS 2021 referees and editors for helpful feedback. I am solely responsible for the findings reported here.

COMPASS '21, June 28–July 02, 2021, Virtual Event, Australia

ACM ISBN 978-1-4503-8453-7/21/06…$15.00
https://doi.org/10.1145/3460112.3471971

pipeline operators utilize sensors to probe oil and gas pipelines for the purpose of localizing and sizing different defect types for maintenance [8].

Less attention has been paid to sociodemographic information to the best of my knowledge, even though households in proximity of such resources may be of special interest to policy makers. For example, political incentives are potentially important [9]. The paper is also in line with recent work in economics that focuses on machine learning as a complement to more common work in causal inference [10, 11]. I close the paper by advocating for greater use of social science in sustainable development by the oil and gas industry and policy makers. Although the focus on standard petroleum studies analyses must continue, social science data may assist to unpack social contexts.

2 SOCIODEMOGRAPHIC AND ECONOMIC DATA FROM ACROSS AFRICA

The analyzed countries are Burkina Faso; Burundi; Democratic Republic of the Congo; Cote d'Ivoire; Cameroon; Ethiopia, Gabon; Ghana; Guinea; Kenya; Liberia; Lesotho; Mali; Malawi; Mozambique; Namibia; Rwanda; Sierra Leone; Senegal; Eswatini; Tanzania; Togo; Uganda; Zambia; and Zimbabwe. Only Lesotho does not border another country in the dataset. The national borders of African nations were historically partitioned without respect to ethnicity, and as such, the data has many households that live close to borders shared with other countries, according to a study that used the same data to analyze HIV/AIDS [12]. Also, it is common for African ethnicities to be found in multiple countries as is the case in these data. The data is from 1997 to 2014.

Table 1 summarizes the data, which consists of surveyed households in pixels, which are located in (or transcend) the above countries. To create the pixels, the African continent is divided up into pixel units of 12.5 km x 12.5 km to access the various variables that provide rich context to the households under study, a common approach [13, 14]. At the household level, the data focuses on gender, with the ethnicity and age of the wife in the household being recorded, as well as whether or not she has not had any access to education or is HIV positive. There are 634 ethnicities represented in the data. At the pixel level, the data also contains light density or luminosity (a common proxy for economic growth [15, 16]), population density, whether or not the area has a body of water, average elevation (in km), soil suitability, whether or not the pixel contains an oil field, whether or not the pixel contains a diamond mine, the malaria stability index as a health proxy, the distance form the pixel centroid to the coast, the distance of the pixel centroid to the capital, the minimum of monthly rainfall, the minimum humidity, whether or not the country's legal system is based on common law or civil law. The modeling results also consists of dummies for the countries and subregion (West, East, Central or Southern Africa). The target to be predicted is whether or not a household is located in a pixel that contains an oil field.

3 DEEP LEARNING

We will present a neural network which learns using the backpropagation algorithm [17] and two hidden layers. An illustration is summarized in Figure 1. Panel A shows a generic neural network [18]. Panel B summarizes the model application. In this neural network model, the neural network is defined as that which takes the input and passes them into the 64-dimension first hidden layer, then takes the output of the first layer and passes them into the 8-dimension second layer; takes the output of the second layer and passes them into the last layer for prediction. The output of the last layer is the prediction. The last layer is activated with a sigmoid function.

The approach uses the *Sequential keras* function as the skeleton of the neural network model, and sequentially add the layers onto the skeleton. After building the layers, we need to compile the model and define the optimizer, loss function and evaluation metrics to optimize our model. In this example, we use the *adam* optimizer to minimize the value of the loss function binary_crossentropy since the goal is to classify a dummy target variable [19]. The quality of the predictions is judged by accuracy.

The neural network predicts the presence of petroleum in ethnic regions with ethnic regions with 89.7% accuracy on the training set and when making predictions on the test set, the accuracy is 89.9%. The loss function suggests that the model has comparable performance on both training and testing datasets. I find it interesting that the sociodemographic variables are able to offer this strong out-of-sample prediction and it is possible that this may be correlated with the physical variables of interest to petroleum science researchers. The model loss is depicted in Figure 2, with both training and test losses being relatively similar, although the learning is relatively slow. The loss function is not convex, perhaps in part due to the model consisting of multiple layers. The increased accuracy in the detection of petroleum fields might be useful in using sociodemographic data to unpack the context of natural resources. At the same time, the slight increased accuracy in the detection of petroleum fields might be useful in using socioeconomic data to unpack the context of natural resources. One possible reason for the correlations may be that demographics may be picking up households who migrated to oil regions for employment or adjacent reasons, although the model is unable to evaluate such a claim.

Although deep learning models tend to be relatively opaque, dimensionality reduction can help visually unpack some of the local and global structure in the data at large. This is done in the next subsection. The approach used is an attempt to accommodate and make more transparent the commonly highly non-linear structure, noise and continuous progressive nature expected in panel datasets.

3.1 Nonlinear Dimensionality Reduction with PHATE

I use nonlinear dimensionality reduction to visualize the data. PHATE involves computing a localized Markov transition matrix (henceforth called a diffusion operator) between units. This operator is computed by first computing local affinities between points then normalizing the affinities such that they become transition probabilities between units. Then we power or diffuse the matrix to obtain longer-range, cleaner connections between units. We then transform these transition probabilities. Finally, we embed the resultant matrix with non-metric multi-dimensional scaling for visualization in low dimensions, which preserves monotone

Table 1: Descriptive Statistics

Variable	Observations	Mean	Std. Dev.	Min	Max
HIV	297000	0.07	0.254	0	1
Wife age	313000	28.763	9.919	15	64
Wife with no education	313000	0.303	0.46	0	1
Ethnicity	290000	453.722	249.627	1	835
Pop. Density	297000	373.423	1197.401	0.111	24373.78
Night lights(economic growth)	297000	3.992	10.661	0	62.978
Land surface area (of each ethnic-country region)	297000	36.139	39.062	0.109	408.988
Mean elevation	297000	0.783	0.608	0	2.181
Soil suitability	297000	0.484	0.224	0.001	0.979
Malaria index	297000	0.687	0.302	0	1
Petroleum	297000	0.066	0.248	0	1
Diamond	297000	0.241	0.428	0	1
Distance to capital	297000	0.386	0.369	0.01	1.882
Distance to sea	297000	0.495	0.365	0.001	1.598
Distance to border	297000	0.115	0.112	0	0.617
GDP per capita	313000	1497.543	1284.933	560.298	16618.26
Absolute latitude	313000	10.538	7.401	0.2	36
Longitude	313000	19.763	17.102	-24.044	57.794
Min. rainfall	313000	8.124	12.79	0	69
Max. humidity	313000	69.943	10.126	35	95
Min. temperature	313000	8.248	6.668	-9	19
West Africa	313000	0.308	0.462	0	1
East Africa	315000	0.336	0.472	0	1
Central Africa	308000	0.143	0.351	0	1
Southern Africa	312000	0.252	0.434	0	1
Common law	304000	0.428	0.495	0	1

[a] Summary statistics of the variables.

Layer type	Observations	Mean
Dense_1 (Dense)	(None, 64)	5376
Dense_2 (Dense)	(None, 8)	520
Dense_3 (Dense)	(None, 1)	9

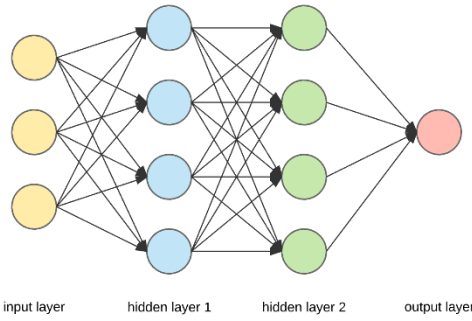

Figure 1: (Panel A) A standard depiction of a neural network with two hidden layers [19].

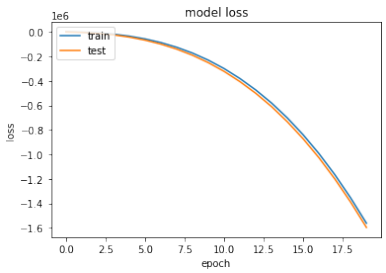

Figure 2: Model Loss: Similarly declining model loss for training and test data. Total parameters: 5,905; Trainable parameters: 5,905; Non-trainable parameters: 0.

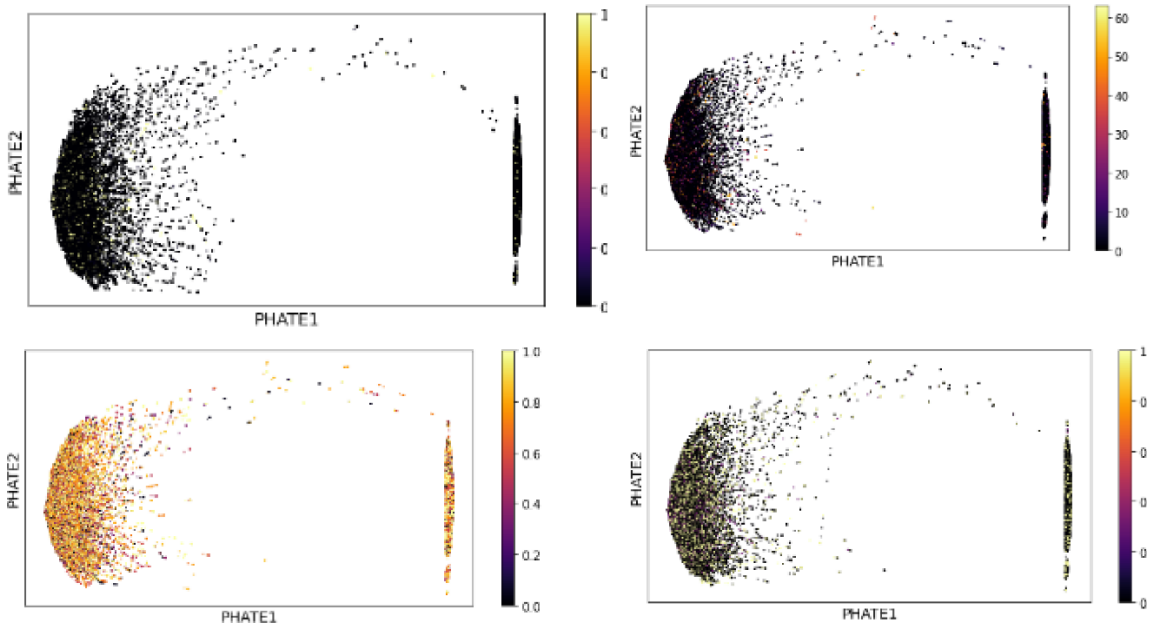

Figure 3: PHATE visualizations with default hyperparameter settings for petroleum (top left); economic growth/night lights (top right); malaria index (lower left); wife in the household with no education (lower right).

relations between potential distances. The discussion is based on Moon et al (2019).

3.2 The PHATE Approach and Algorithm

As shown in the empirics of the section, the choice of a kernel K and bandwidthε affect the presentation of the results. In the previously-discussed methods, the bandwidth choice represents a tradeoff between encoding local and global information in the probability matrix p_ε. If ε is too small, the neighbors of the points in sparsely sampled regions on the African continent may be excluded entirely and the trajectory structure in p_ε will not be encoded. However, if ε is too big, then p_ε loses local information as $P_\varepsilon(x)$ becomes uniform for all x. PHATE uses an adaptive bandwidth that changes with each point to be equal to its k-th nearest neighbor as well as an α-decaying kernel that controls the rate of decay of the kernel. The trajectory structure yielded by the PHATE approach is not artificially generated, but dominant in the dataset. Therefore, the PHATE visualization will only show trajectory structures when data fits such a geometry; otherwise, other patterns of clustering will be expressed in the PHATE visualization. The algorithm follows [6].

4 PHATE: PETROLEUM FIELDS AND SOCIODEMOGRAPHIC DATA VISUALIZATIONS

Figure 3 summarizes the data structure while focusing on a selection of the variables. Here, I first use the default KNN or number of nearest neighbors of 5, and an alpha decay of 15. The t or number of times the operator is powered is held to the default automatic selection. Figure 3 shows the data while zooming in on the petroleum variable, as well as a selection of social and demographic variables. In all cases, the data is reduced to two dimensions, PHATE 1 and PHATE 2. Figure 3a zooms in on the presence of petroleum fields although these are not very visible. Figure 3b focuses on economic growth using satellite night light data, and even though the structure is preserved, these are also relatively low (the mean luminosity representing growth is about 3.99 units, ranging from 0 to 63. Figure 3c shows malaria prevalence at the pixel level, and these are relatively high in the data. A simple OLS regression shows a negative 9% correlation between malaria prevalence and the presence of a petroleum field (significant at the 0.001% level), which suggests that the neural network might have picked up the negative correlation. This is intuitive to the extent that oil firms may select on regions with highly productive workers. Figure 4c shows the presence of wives in households with no education. While not as common as the malaria prevalence in Figure 3, it still appears to be high. A simple OLS regression shows a negative 1% correlation between malaria prevalence and the presence of a petroleum field (significant at the 0.001% level). Given that malaria has such outlier high prevalence in its PHATE diagram, we revisit the hyperparameter filtering thresholds for closer examination.

In Figure 4, I change the parameters, setting the KNN to 4, the decay to 15 and t is now 12. In all cases, the households are presented as more dispersed across the two dimensions, although the clustering to the left mostly remains intact. Both the local and global structures remain quite consistent.

I vary the parameters further, setting the KNN to 40, keeping the decay at 15 and changing t to 60. In all cases, the households are presented as even more dispersed across the two dimensions than in the previous presentations. The structure or overall shape does not appear to be significantly different from the previous presentations.

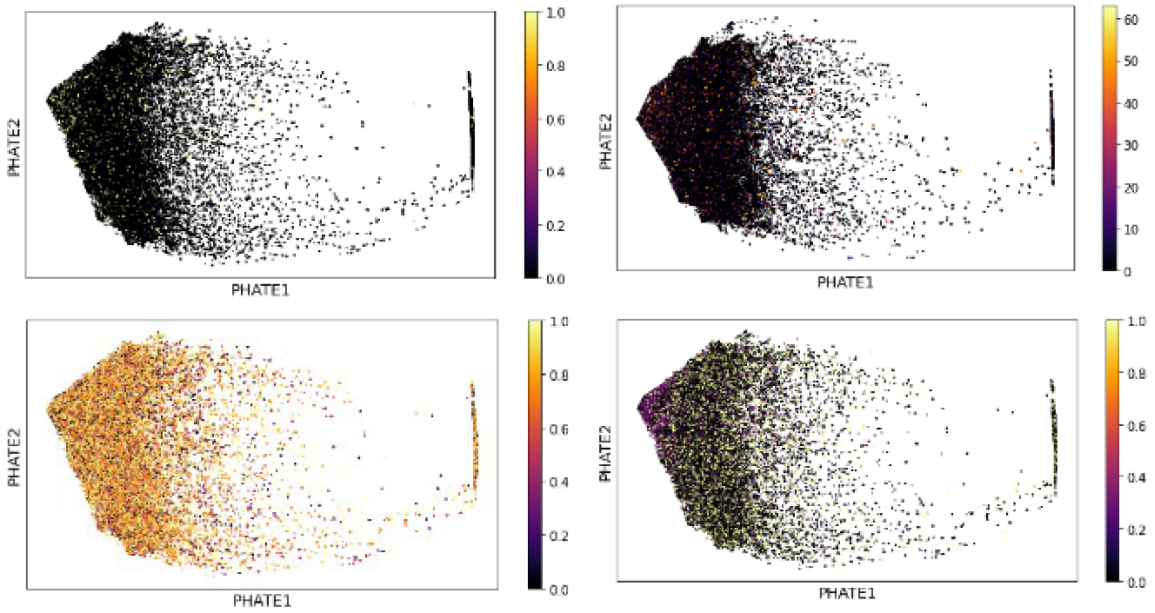

Figure 4: PHATE visualizations with slightly-changed hyperparameter settings for petroleum (top left); economic growth/night lights (top right); malaria index (lower left); wife in the household with no education (lower right).

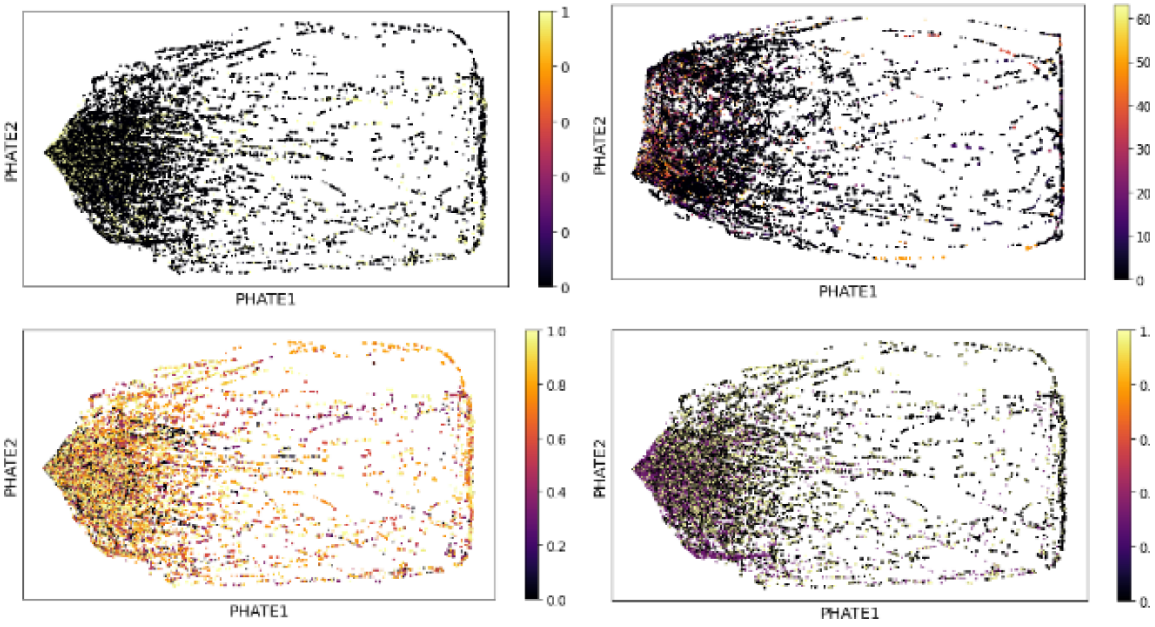

Figure 5: PHATE visualizations with significantly-changed hyperparameter settings for petroleum (top left); economic growth/night lights (top right); malaria index (lower left); wife in the household with no education (lower right).

In just about all of the cases, the sociodemographic data visualization patterns appear similar to the petroleum data. This is pattern tends to be consistent even as one varies the PHATE parameters across different variables, although the difference between Figure 5 and Figure 3 are relatively stark. The socioeconomic variable that has features the most in the PHATE visualizations is consistently the malaria prevalence variable, which has the highest outcomes across the board. Although there are some regions of the plot with higher variable outcomes than others, these household observations are fairly well distributed over the plot. PHATE is efficient at preserving global structure, and increasing its hyperparameters generally keeps its inter-cluster relations meaningful when different variables

are focused on, although it still warps the high-dimensional shape of the data.

5 CONCLUSION

For sustainable development, oil exploration needs a relatively inclusive approach. Much research effort has shown that inclusive institutions have positive effects [20, 21], and this may also hold for institutions focused on oil resources. This paper offers a possible initial step in this discussion: the possibility of using socioeconomic data to unpack the presence of oil resources. Such an approach may complement qualitative surveys about experiences with oil resources. The analysis focused on oil resources already discovered, not unknown fields (which would have a different data-generating process). In future work, firms may wish to combine social and engineering data for newer explorations, while ensuring that ethics and social impact are adhered to in African communities.

REFERENCES

[1] Halvor Mehlum, Karl Moene, and Ragnar Torvik 2006. Institutions and the resource curse. Economic Journal, 116(508), 1–20.

[2] Daron Acemoglu, Amy Finkelstein, and Matthew J. Notowidigdo 2013. Income and health spending: Evidence from oil price shocks. *Review of Economics and Statistics*, 95(4), 1079–1095.

[3] F. van der Ploeg. 2011. Natural resources: curse or blessing? *Journal of Economic Literature*, 49(2), 366–420.

[4] van der Maaten, L. and G. Hinton (2008). "Visualizing data using t-SNE." *Journal of Machine Learning Research*, 9, 2579–2605.

[5] McInnes, L., J. Healy, and J. Melville (2018). "UMAP: Uniform manifold approximation and projection for dimension reduction." ArXiv Preprint arXiv:1802.03426.

[6] Moon, K.R., D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim,, A. van den Elzen, M. J. Hirn, R. R. Coifman, and N. B. Ivanova, (2019). "Visualizing structure and transitions in high-dimensional biological data." *Nature Biotechnology*, 37(12), 1482–1492.

[7] Alvarado, V., A. Ranson, K. Hernandez, E. Manrique, J. Matheus, T. Liscano, and N. Prosperi (2002, January). Selection of EOR/IOR opportunities based on machine learning. *In European Petroleum Conference*. Society of Petroleum Engineers.

[8] Mohamed, Abduljalil, Mohamed Salah Hamdi, and Sofiène Tahar (2015). "A machine learning approach for big data in oil and gas pipelines." *2015 3rd International Conference on Future Internet of Things and Cloud*. Institute of Electric and Electronic Engineers.

[9] Robinson, James A., Ragnar Torvik, and Thierry Verdier (2006). "Political foundations of the resource curse." *Journal of Development Economics*, 79(2), 447–468.

[10] Mullainathan, Sendhil, and Jann Spiess (2017). "Machine learning: an applied econometric approach." *Journal of Economic Perspectives*, 31(2), 87–106.

[11] Storm, Hugo, Kathy Baylis, and Thomas Heckelei (2020). "Machine learning in agricultural and applied economics." *European Review of Agricultural Economics*, 47(3), 849–892.

[12] Anderson, Siwan (2018). "Legal origins and female HIV." *American Economic Review*, 108(6), 1407–1439.

[13] Michalopoulos, S., and E. Papaioannou (2013). "Pre-colonial ethnic institutions and contemporary African development." *Econometrica*, 81(1), 113–152.

[14] Michalopoulos, S., and E. Papaioannou (2014). "National institutions and subnational development in Africa." *Quarterly Journal of Economics*, 129(1), 151–213.

[15] Henderson, Vernon, Adam Storeygard, and David N. Weil (2011). "A bright idea for measuring economic growth." *American Economic Review*, 101(3), 194–199.

[16] Henderson, J. Vernon, Adam Storeygard, and David N. Weil (2012). "Measuring economic growth from outer space." *American Economic Review*, 102(2), 994–1028.

[17] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors." *Nature*, 323(6088), 533–536.

[18] Ognjanovski, G. (2019). "Everything you need to know about Neural Networks and Backpropagation." https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a

[19] Chollet, Francois (2015). *Keras*. https://keras.io.

[20] Acemoglu, D., and J. A. Robinson, (2019). "Rents and economic development: the perspective of Why Nations Fail." *Public Choice*, 181(1), 13–28.

[21] Roy, Ananya, Genevieve Negrón-Gonzales, Kweku Opoku-Agyemang, and Clare Talwalker. *Encountering poverty: Thinking and Acting in an Unequal world*. Vol. 2. Univ of California Press, 2016.