Model to identify UK industries potentially susceptible to employment shocks due to automation

Abstract

Expected outcome of this paper is to formalise a model that can identify a sector with potential decrease of the employment due to automation within the next 3-5 years. Model is based on the assumption that if employment variations in a given industry or sector are highly correlated with much more obvious economic indicators such as negative economic growth or fluctuations in the output of this particular industry, there is a significantly smaller chance that the new disruptive technology is the cause of these variations, and therefore employment in this industry has a smaller probability of being negatively affected in the short-term and vice versa.

Introduction

Emergence of automation is not a novel concept. Economic theories of Adam Smith and Karl Marx were heavily influenced by the idea, that the machinery has significant influence on economic performance and employment. (Smith, 1776) , (Marx, 1973).

Recent rise in the adoption of artificial intelligence technologies in manufacturing, retail, logistics have caused many economists and sociologists to propose alternative social and economic policies to facilitate the adaptation of the society to the inevitable, yet uncertain effects of automation. These speculative policies range from basic income guarantee (Ford, 2015) to embracing the full automation and shortening workweek. (Srnicek and Williams, 2015)

I will examine four industries in the UK: retail, administration, scientific research, and manufacturing. Retail, administration and manufacturing are amongst " the minor occupation groups, or three-digit occupations, with the lowest probabilities of future increased demand" and scientific jobs are amongst groups with "the greatest probabilities of future increased demand" in the UK, according to "Future

of Skills: Employment in 2030" (*FOS*) report (Bakhshi, Downing, Osborne, Schneider; 2017)

The model labels each examined industry as susceptible or not susceptible to automation in the future short-term period of three to five years.

I will then compare the findings of the model with the data presented in the *FOS* report.

Data

I have used Gross Value Added, FTSE100 and Research & Development expenditure as independent variables for three main reasons.

1) **Availability**

   While it is not always easy to find data directly related to automation, the discussed datasets are easily accessible for general public.

2) **Verifiability**

   In addition to being easily available, the data presented in the model is simple to verify.

3) **Ubiquity**

   The variation of these datasets can be found in most developed and many developing countries. The acknowledged shortcoming is that GVA and R&D information is usually only available for large industries and omits the smaller segments of the economy.

Model

Expected Outcomes

The model determines how well the chosen variables predict the changes in the number of workers in a particular industry and measures the difference between this prediction and the actual data. Based on this difference it classifies the industry as susceptible or not susceptible to negative short-term employment

shocks due to automation.

These classifications should not be regarded as definite or conclusive, nor should they be used to assess the probabilities of susceptibility of a particular industry or probabilistic differences between multiple industries. Statements such as, industry X is 75% susceptible to automation, or industry X is 25% more likely to be susceptible to automation than industry Y, have no meaning and should not be made.

It is necessary to emphasise that the model does not search for the best set of methods and variables that predict the employment figures, but rather evaluates how well given independent variables describe the actual data. Therefore the initial selection of the variables defines the outcome of this model.

Key characteristics

1) **Accessibility**

   All the data used for this model is publicly available.

2) **Usability**

   Model focuses on short-term predictions.

3) **Dynamics**

   This model does not require excessive amounts of historical data.

Key assumptions

1) **Transitivity of preference (Regenwetter and Dana, 2011) for chosen variables**

   In terms of the interpretability, the model always prefers the three chosen independent variables to other variables, although there is no preference between these three variables. For any variable outside the list of the chosen independent variables X,Y,Z model strictly prefers the ones with the highest correlation to dependent variable. Let's assume that there exists a variable V that is highly correlated to the dependent variable. Then, if V is

highly correlated to one of the chosen variables, by property of multicollinearity we only keep X,Y,Z and leave V out. If V is not correlated to given independent variables, this difference should be reflected in the coefficient described below.

2) **Omission of multicollinearity among chosen variables**

Even if the chosen variables are highly correlated among each other, we do not exclude any of these variables from the model in order not to violate the previous assumption.

3) **Causation between automation and employment**

The model assumes that automation influences employment to at least some degree, if the given independent variables are unable to explain the changes in employment.

Methodology

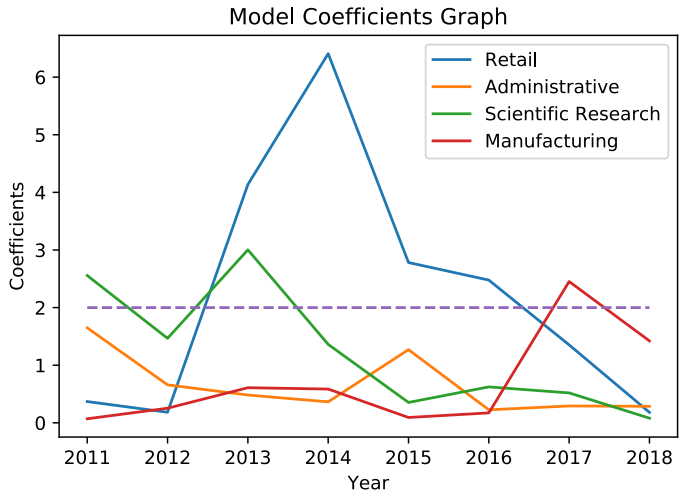
The model uses multivariable linear regression, one of supervised machine learning methods. It takes the values for both independent and dependent variables for the previous five years as training data. Subsequently it predicts the value of the dependent variable for the sixth year. Then, the two standard deviation values are calculated. In order to calculate the first standard deviation we take five actual values for dependent variable for the previous five years, let's call it STD_{actual}. For the second standard deviation value we take the predicted and the actual values of the dependent variables for the sixth year, let's call it STD_{model}. The necessary coefficient is

$$Coeff = \frac{STD_{model}}{STD_{actual}}$$

The nominal difference between the actual and predicted values is also calculated. If the nominal difference is negative and the model coefficient is greater than 2 for more than two years in a row, the industry is classified as susceptible to negative employment shocks due to automation. (See Figures 1 and 2 for coefficients calculated for different years.)

Figure 1

Year	Actual	Predicted	STD Actual	Nominal Difference	STD Model	Model Coefficient
Retail						
2011	4,045,487.75	4,005,472.87	76,558.75	- 40,014.88	28,294.79	0.37
2012	4,088,587.75	4,071,088.87	67,401.61	- 17,498.88	12,373.58	0.18
2013	4,037,178.25	4,400,004.86	62,043.62	362,826.61	256,557.15	4.14
2014	4,039,421.50	3,461,566.37	63,767.93	- 577,855.13	408,605.28	6.41
2015	4,091,392.50	4,017,077.24	18,887.93	- 74,315.26	52,548.83	2.78
2016	4,194,233.25	4,109,824.26	24,091.54	- 84,408.99	59,686.17	2.48
2017	4,218,465.75	4,114,024.83	54,589.22	- 104,440.92	73,850.88	1.35
Administration						
2011	1,328,130.25	1,402,549.43	31,914.01	74,419.18	52,622.31	1.65
2012	1,368,528.50	1,398,273.33	31,912.83	29,744.83	21,032.77	0.66
2013	1,416,163.00	1,397,152.09	27,848.84	- 19,010.91	13,442.74	0.48
2014	1,441,981.75	1,460,924.86	36,767.50	18,943.11	13,394.80	0.36
2015	1,504,064.25	1,418,860.21	47,425.43	- 85,204.04	60,248.35	1.27
2016	1,515,261.50	1,495,730.09	61,199.73	- 19,531.41	13,810.80	0.23
2017	1,576,542.00	1,548,639.83	67,380.15	- 27,902.17	19,729.81	0.29
Science						
2011	1,873,349.00	2,054,339.84	50,056.36	180,990.84	127,979.85	2.56
2012	1,949,692.00	1,886,783.36	30,331.37	- 62,908.64	44,483.13	1.47
2013	2,048,205.00	1,929,230.80	28,014.81	- 118,974.20	84,127.46	3.00
2014	2,146,108.75	2,026,201.23	62,266.57	- 119,907.52	84,787.42	1.36
2015	2,177,891.25	2,229,624.88	103,348.56	51,733.63	36,581.20	0.35
2016	2,284,069.25	2,177,212.43	121,152.87	- 106,856.82	75,559.18	0.62
2017	2,342,070.00	2,444,336.32	139,219.84	102,266.32	72,313.21	0.52
Manufacturing						
2011	2,863,214.00	2,843,922.74	193,834.97	- 19,291.26	13,640.98	0.07
2012	2,903,987.00	2,838,193.39	183,310.05	- 65,793.61	46,523.11	0.25
2013	2,924,962.00	2,794,957.45	150,780.95	- 130,004.55	91,927.10	0.61
2014	3,009,640.00	2,942,277.11	81,245.79	- 67,362.89	47,632.75	0.59
2015	2,995,573.00	3,003,660.69	61,508.88	8,087.69	5,718.86	0.09
2016	3,000,646.00	2,986,773.52	56,804.01	- 13,872.48	9,809.32	0.17
2017	2,930,773.00	3,123,235.49	55,488.89	192,462.49	136,091.54	2.45

Figure 

2

Limitations

1) Handling the outliers

The two central statistical methods used are Pearson correlation coefficient and linear regression model. The biggest limitation of both of these methods is the inadequate ability to deal with the outliers. The attempts to mitigate this limitation were made by plotting time-series graphs and examining for inconsistencies between visual and numerical representation of the studied data. (See figure 3)

Figure 3

2) Incomplete interpretability

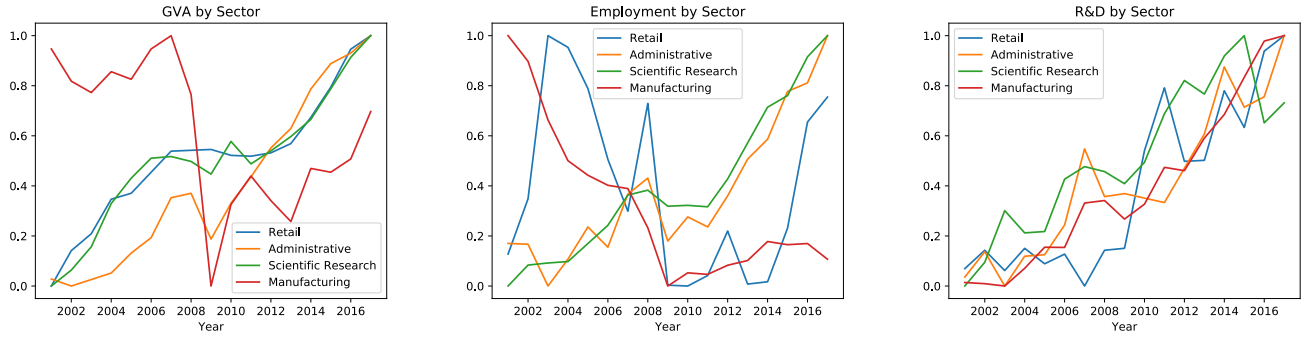
The model only gives the classification lables, it can not however confidently predict whether the negative shocks are due to automation or some other factor. Therefore, the model heavily relies on the assumption number three described above.

Results

The analysis of data has shown that three independent variables described above explain most of the fluctuations in the employment figures for science and administration industries, for the period 2001-2017. Testing for relationship between each independent variable and dependent variable separately, showed set of robust correlations. Simple linear regression model explains between 65% and 90% of these relationships. Derived model has been able to predict the employment figures rather accurately, so these industries are labeled as not susceptible to employment changes in the short term due to automation or technology disruption.

For manufacturing industry, GVA to employment and R&D expenditure to employment correlation graphs are almost mirror reflections of one another. Linear regression model does not explain the correlations as well as in the previous two industries but R-squared is still between 0.45 and 0.54. Therefore,

Normalised Graphs



we still can infer that a big part of employment changes in manufacturing industry

is due to the more obvious reasons. Proposed model labels manufacturing as not susceptible to automation in the short term, although strong negative correlation between R&D expenditure and employment, does make the labeling less conclusive.

All three indicators poorly predict number of people employed in the retail industry. The model used labels this industry as susceptible to negative employment changes due to automation.

Therefore, model short-term predictions confirm findings presented in *FOS* report about retail and scientific research industries but contradict the ones about administration and manufacturing sectors.

Figure 4

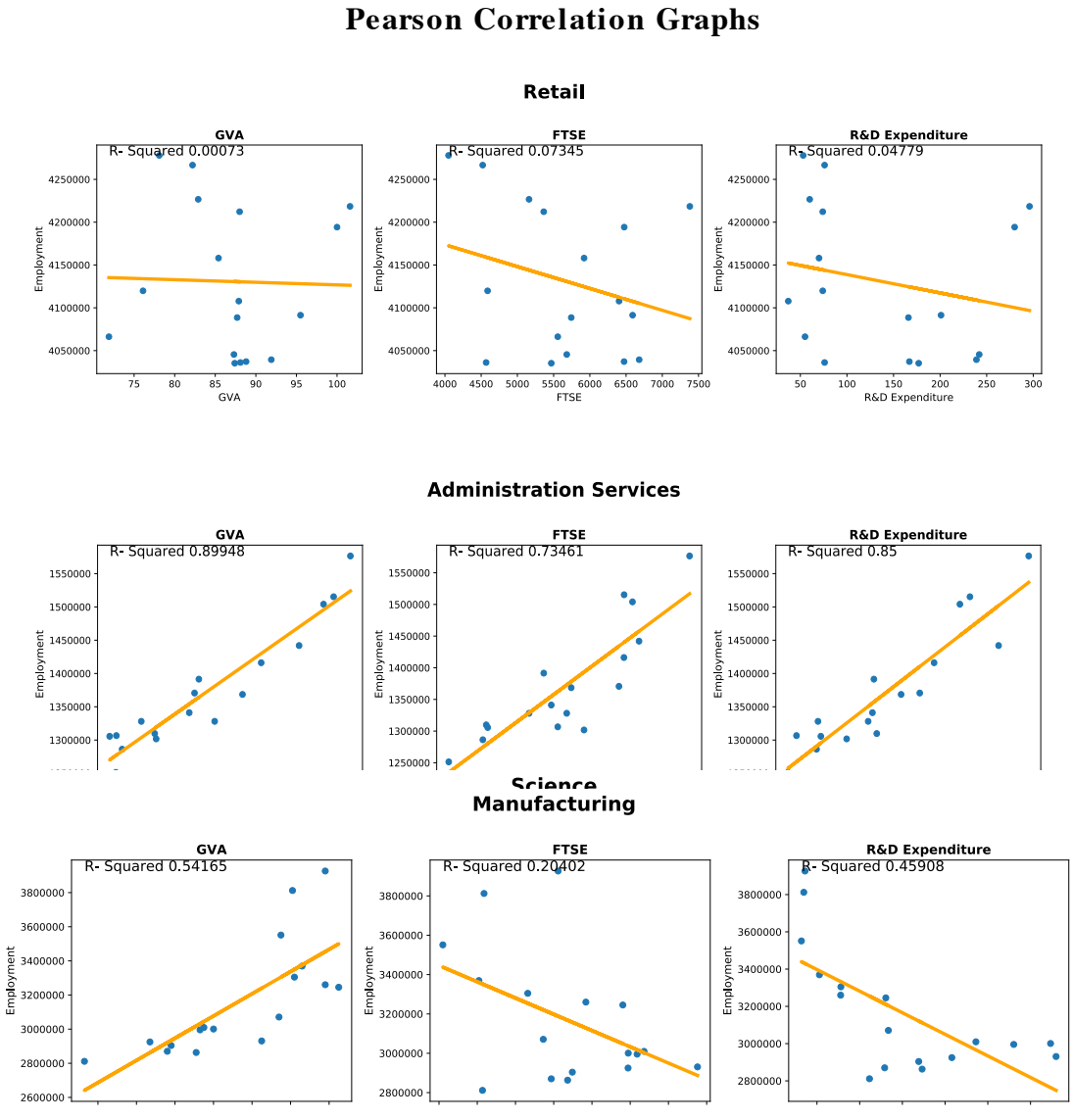


Figure 5

Correlation Tables By Sector 2001-2017

Retail

	GVA	FTSE	RD Expenditure	Employment
GVA	1	0.7927	0.2476	**0.4445**
FTSE	0.7927	1	0.2160	**0.0935**
RD Expenditure	0.2476	0.2160	1	**-0.2313**

Science

	GVA	FTSE	RD Expenditure	Employment
GVA	1	0.7515	0.6851	**0.8871**
FTSE	0.7515	1	0.8155	**0.9568**
RD Expenditure	0.6851	0.8155	1	**0.8713**

Admin

	GVA	FTSE	RD Expenditure	Employment
GVA	1	0.9240	0.9023	**0.9584**
FTSE	0.9240	1	0.9405	**0.8959**
RD Expenditure	0.9023	0.9405	1	**0.8579**

Manufacturing

	GVA	FTSE	RD Expenditure	Employment
GVA	1	-0.0928	-0.4218	**0.7360**
FTSE	-0.0928	1	0.8347	**-0.4517**
RD Expenditure	-0.4218	0.8347	1	**-0.6776**

Conclusion

This model does not aim to give an in-depth understanding of the properties of any particular technology, evaluate specific probabilities of the emergence of automation or calculate the effect of the damage technology can bring to the existing labour force, but rather to help users of this model to identify industries where the data is incongruent, so they can investigate the issue further.

This model should be tested on a significantly higher number of industries before we can conclude that it can be extrapolated to a wider scope of industries or

geographies. Furthermore research methods such as logistic regression and more in-depth study of data distribution patterns can increase the accuracy of the model. Also, there is a potential to build an extended machine learning classification model by having a larger set of data points in order to facilitate the identification of the industries susceptible to automation.

References

1. Smith A, (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, pp.277
2. Marx K, (1973). *Grundrisse*. Penguin Books in association with New Left Review, pp.615
3. Ford M, (2015). *Rise of the Robots*. Basic Books, pp.257
4. Srnicek N, Williams A (2015). *Inventing the Future: Postcapitalism and a World Without Work*. Verso. Kindle edition, loc.2300
5. Bakhshi H, Downing J, Osborne M, Schneider P. (2017). *Future of Skills: Employment in 2030*. Creative Commons. pp. 49-50. Available from: *https://futureskills.pearson.com/research/assets/pdfs/technical-report.pdf*
6. Regenwetter M, Dana J(2011). *Transitivity of Preferences*. American Psychological Association, Vol. 118, No.1, 42-56.

Data

1. Gross Value Added. Available from: https://www.ons.gov.uk/economy/grossvalueaddedgva/datasets/nominalandrealregionalgrossvalueaddedbalancedbyindustry
2. Research and Development Expenditure. Available from: https://www.ons.gov.uk/economy/governmentpublicsectorandtaxes/researchanddevelopmentexpenditure/datasets/businessenterpriseresearchanddevelopmenttimeseriesspreadsheet
3. FTSE100 index. Available from: https://www.londonstockexchange.com/statistics/ftse/ftse.htm