MINING ACADEMIC PUBLICATIONS TO PREDICT AUTOMATION

Dartmouth Computer Science Technical Report TR2020-892

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Bachelor of Arts

in

Computer Science

by

Elena Doty

Advisor: Soroush Vosoughi

DARTMOUTH COLLEGE

Hanover, New Hampshire

June 2020

Abstract

This paper proposes a novel framework of predicting future technological change. Using abstracts of academic publications available in the Microsoft Academic graph, co-occurrence matrices are generated to indicate how often occupation and technological terms are referenced together. This matrices are used in linear regression models to predict future co-occurrence of occupations and technologies with a relatively high degree of accuracy as measured through the mean squared error of the models. While this work is unable to link the co-occurrences found in academic publications to automation in the labor force due to a dearth of automation data, future work conducted when such data is available could apply a similar approach with the aim of predicting automation from trends in academic research and publications.

Contents

Chapter 1

Introduction

Automation is a growing trend in many industries, as technology develops to the point of being able to do many manufacturing and service jobs more proficient than humans. With this trend comes increasing worry and uncertainty, especially among blue collar and lower wage workers, over job security and future skill requirements to acquire and hold a job. The ability to predict job areas that are likely to be overcome by automation and the image of future occupational requirements would help social planners and individuals to prepare for future changes.

Academic research is often years ahead of industry and society in terms of technological process and developments, and as such can provide a sense of future trends in industry and society. Past research has shown that community reactions to new publications, as measured by citation and download counts, can be predicted through citation content, text content, or both. Similar models have been successful in predicting areas of future research advancement.

This work attempts to predict trends in automation and job skill requirements from current publications. The ability to predict these attributes of the labor market would allow academic and other institutions to better prepare their students for success in future careers and industries. A sense of which jobs are likely to be extinct would help programs tailor their offerings toward the requirements of the future, while an idea of which skills will be in high demand would better guide students' academic plans and choices. Accordingly, the models proposed above could be very useful to many groups should they prove successful, even if they are limited to broad predictions of future trends.

Section 1.1

Microsoft Academic Graph

The Microsoft Academic Graph (MAG) is an automatically assembled graph of more than 120 million academic papers by more than 114 million authors. Entries in the graph may contain information on the paper's author, venue, and references, as well as an indexed abstract. The graph is skewed toward papers published after 1960, though some entries date back to the 1800s. Herrmannova and Knotch (2016) find that citation data and publication metadata in the MAG correlate well with external databases, and that the MAG, while slightly biased toward technical disciplines, has good coverage across fields of study. However, they note that only around 30 million of 127 million papers include citation data and only around 22 million contain affiliate information, suggesting a limitation of completeness.

The MAG is available through Microsoft Azure APIs, but can also be downloaded through Open Academic Graph as compressed JSON files.

Section 1.2

Related Literature

As automation is a topic of concern to many workers, politicians, and companies, many studies have been done attempting to predict the impact of automation and the risk of automation in the future. A prominent work in the field is one study done by Carl Frey and Michael Osborne, “The future of employment: How susceptible are jobs to computerisation?”. They develop a process to determine the susceptibility of jobs to automation that considers each occupation as a collection of tasks. Each task’s automation potential is assessed, and use these assessments to predict a probability that the job itself is computerisable and categorize it into low, medium, or high automation potential. They estimate that 47% of American jobs are considered highly automatable, with employment in services occupations especially at risk.

In response to Frey and Osborne, many others have suggested that the actual percent of susceptible jobs is lower. Arntz, Gregory, and Zierahn (2017) argue that accounting for the heterogeneity of jobs decreases the share to only 9%. They propose that a job-level approach, rather than an occupation-level approach, decreases the high polarity of assignment (where most jobs were assigned either the low or high categories of risk) and such that only 9% of jobs face an automation potential higher than 70%. Nedelkoska and Quintini (2018) follow a similar approach as Frey and Obsorne, but use the Survey of Adult Skills and more disaggregated occupational classification and estimate for the 32 OECD countries. They find that 14% of jobs in OECD countries participating in the Survey of Adult Skills are highly susceptible to automation (have a probability of automation that is above 70%) and that an additional 30% of jobs have a probability of automation between 50 and 70%. They note large variations across countries, and attribute this to differences in the ways countries organize the same economic sectors.

Other studies have used publication graphs as bases for prediction. Luo, Valenzuela-Escárcega, et al (2018) use a graph of influence statements and citations to find white spaces in scientific knowledge and predict influence links likely to be discovered in the near future. Hahn-Powell, Valenzuela-Escárcega, and Surdeanu (2017) construct a graph of influence relations from a collection of publications and develop a search system to reduce the difficulty of finding academic knowledge in the vast collection of literature available. Uzzi1, Mukherjee1, et al (2013) analyze almost 18 million papers to analyze trends in the highest impact papers (in terms of innovativeness and impact), and find that papers with “exceptionally conventional combinations of prior work” with “an intrusion of unusual combinations” were twice as likely to be highly cited.

To the best of my knowledge, no work has been done attempting to predict automation from publications. This work adds to the literature by introducing a novel way of assessing the future impact of academic work on the labor market and jobs.

Chapter 2

Methods

A randomly selected subset of the papers available in the MAG were reconstructed from the available JSON files and used for analysis. More than 850,000 papers were included in the analysis, with publication dates ranging from 1960 to 2019. Papers for which citation data or abstract were unavailable were excluded from the sample, as were papers in languages other than English. Papers published after 2015 were also excluded to avoid biased results from the increasingly variable number of papers per year included in the MAG. Figure 2.1 presents the distribution of years of publication.

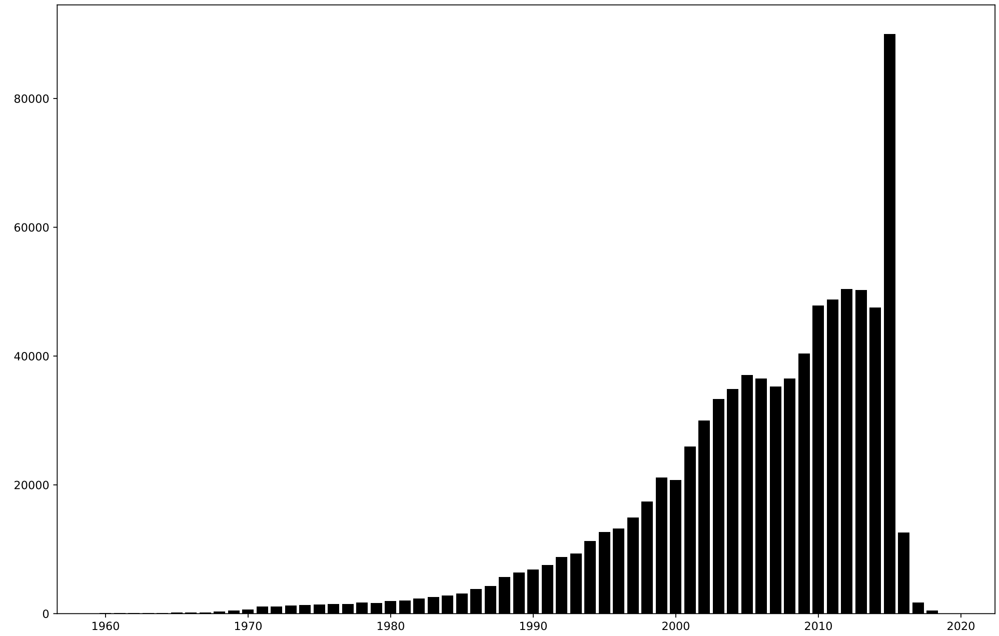

Figure 2.1: Distribution of Year of Publication

Abstracts for each paper were reconstructed from the indexed abstracts available, and were parsed to tag any words representing occupations and terms related to automation or artificial intelligence. These terms were those included in one of two lists: a list of occupations, obtained from the Bureau of Labor Statistics, and a list of terms related to job automation were formed, constructed from a combination of a Wikipedia list of artificial intelligence terms and a list of terms obtained from expansion of a seed list, shown in Table 1, using the Google News Word2Vec model (Mikolov et al, 2013) to obtain the most similar words; more than a thousand technological terms were generated. Before use, this Word2Vec model was fine-tuned using the reconstructed paper abstracts in order to improve the sensitivity of the model to the terms of interest in such a context. Following this training, tagging of occupational and technological terms was done to obtain counts of the co-occurrences of each pair of occupation and term. This procedure was performed multiple times to count co-occurrence overall and in one and five year increments, and the counts for each pair of job and technology terms were combined to create a count matrix for each respective period.

artificial intelligence	automation	algorithm
automate	automatic	automatically
neural network	cyborg	AI
droid	deep learning	automated
computational	machine learning	robot
robotic	bot	optimization

Table 2.1: Seed Words

Figure 2.2 presents a cluster-map generated from the co-occurrence matrix containing counts across all years. The log of count values are used to avoid extreme values from blowing out the map colors. This map suggests that technological research is concentrated in a few industries, and that most sectors have little to no technological research.

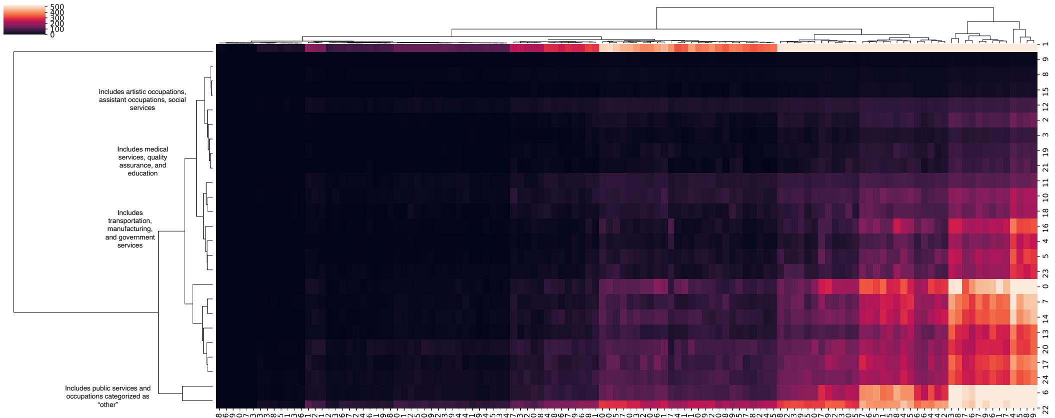

Figure 2.2: Cluster-Map, All Years

Section 2.1

Linear Regression Models

Count matrices were also constructed on a one year basis for use in predictive modeling. These matrices start in 1958 and continue through 2017. A linear regression model was used to predict future count matrices. This model is of the form $y = m_1x_1 + m_2x_2 + ... + m_nx_n$, where y represents the predicted number of co-occurrences between an occupation and technological term in a given year, each x_i represents a different year's co-occurrence count for a given occupation and technological term pair, and each m_i represents the weight or importance assigned to this term by the model. Depending on the number of years used to train the model, n

can range from one to more than sixty. Two different prediction methods were used to assess the potential of the model. First, a single year (2016) was used as the target prediction matrix, and all other years were used to train the model before attempting a prediction and comparing the result to the actual counts for 2016. Second, predictions for each year from 1959 to 2016 were made using the counts for all years before them. Two different models were considered, one which considered only the count for a particular job and term pair and one that considered the counts from similar jobs and terms as part of the predictive vector. Jobs and terms were considered "similar" if they were clustered together by the Word2Vec model described above. These linear regression models were evaluated using mean squared error (MSE) as the main measure of accuracy.

The linear regression model was used to explore several predictive areas. First, each year of data was predicted using all years that came before it to train the model. For example, the values for 2015 were predicted using all values from 1958 to 2014, the values for 2000 were predicted using all values from 1958 to 1999, and the values for 1959 were predicted using only the values from 1958. Predicted values were not used in future predictions. Second, the model was trained on data from 1958 to 2004, then used to predict values for 2005. These predicted values were added to the training data and the model was then used to predict values for 2006. This was done for all years between 2005 and 2015, using each year's prediction as part of the training data for the following years. Finally, the predicted values for each year were compared to the actual data gathered for those years. Mean squared error (MSE) was used to measure performance, and was calculated based on the following formula: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{y_i})$, where n represents the number of observations estimated by the model, y_i represents the true value for the observation, and $\widetilde{y_i}$ represents the model's predicted value for the observation.

To aid in visualization, occupations were clustered into twenty-five clusters using the same Word2Vec model that was used to generate terms and to find similar jobs and terms for use in the second linear model. Hierarchical agglomerative clustering was used throughout this work for all clustering applications. This clustering algorithm works in a bottom-up way, treating each observation as in individual cluster at the outset then agglomerating pairs of clusters successively until there is one cluster that contains all observations.

Chapter 3

Results

Figures 3.1 and 3.2 present the performance of the linear regression model in terms of mean squared error as a function of the number of years used in prediction. Mean squared error (MSE) is measured in units of co-occurrence between each job and technology term pair. Years used in prediction refers to how many years were included in the data used to train the model: one year implies that data from 1958 was used to predict the counts in 1959, two years implies that the data from 1958 and 1959 was used to predict the counts in 1960, and so on. Fifty-seven years implies that data from 1958 to 2015 was used to predict 2015 counts. Figure 3.3 presents a baseline model for comparison, in which only one prior year was used to predict values for each given year.

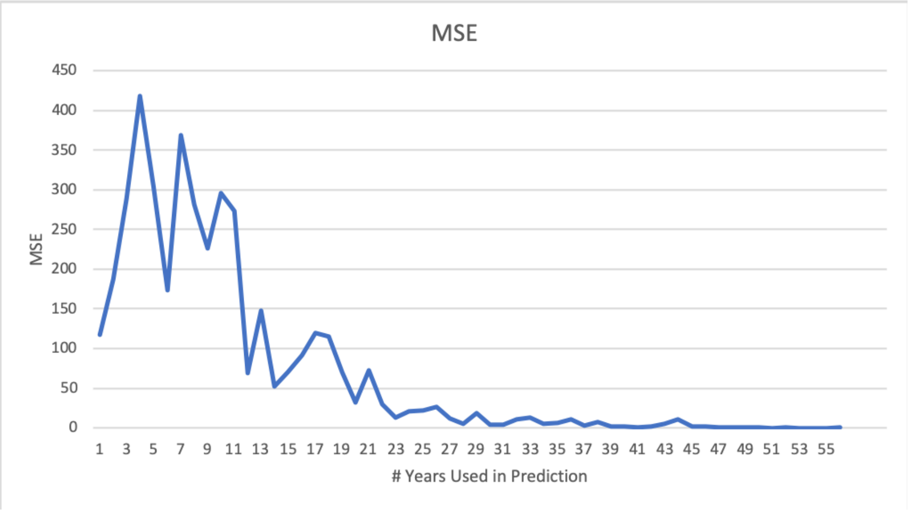

Figure 3.1: Linear model. Predictions based on data from all previous year.

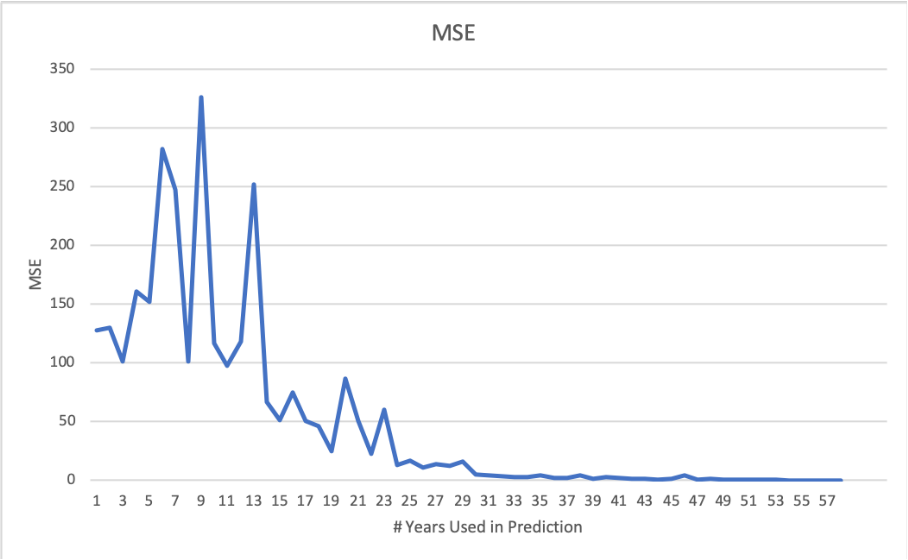

Figure 3.2: Linear model, using similar jobs and terms. Predictions based on data from all previous years and data for similar jobs and terms.

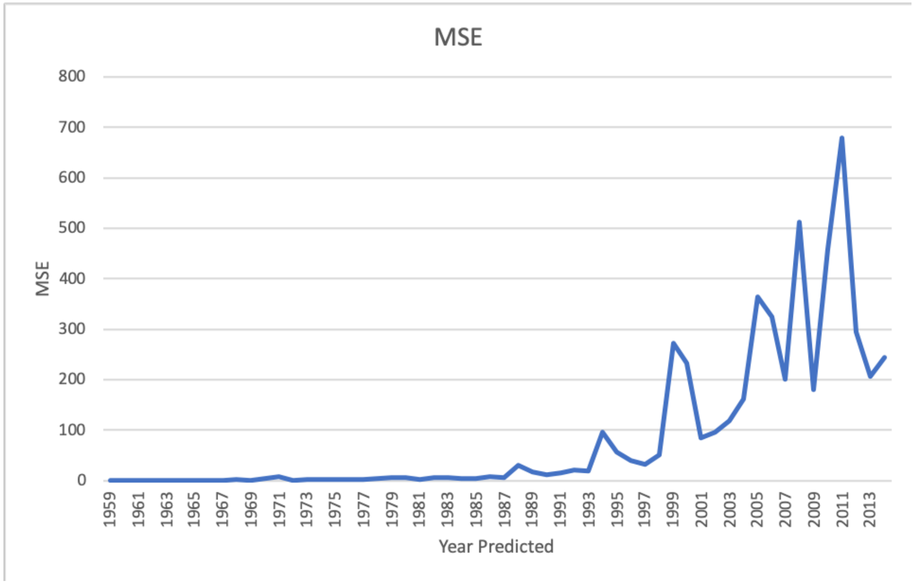

Figure 3.3: Linear model, using similar jobs and terms. Predictions based solely on data from the previous year.

Notably, the model including similar jobs and terms as part of the predictive vector improves the accuracy of the model's predictions.

The model trained on data from 1958 to 2004 was also used to predict count matrices for the years from 2005 to 2015 to evaluate the accuracy of the model's predictions compared to actually data. Heatmaps were constructed from these predicted matrices, and are presented in figures 3.4 through 3.7. Figure 3.4 presents the heatmap generated from predicted data about 2005. For comparison, figure 3.5 presents the heatmap generated from actual data from 2005. In figures 3.6 and 3.7, the same comparison is presented for 2015. On the vertical axis, clusters of occupations are enumerated. On the horizontal axis, unclustered terms are enumerated. Values reported indicate the natural log of the number of co-occurrences between occupation clusters and terms in publications of the year. Values for jobs in the same cluster are summed before the natural log is taken.

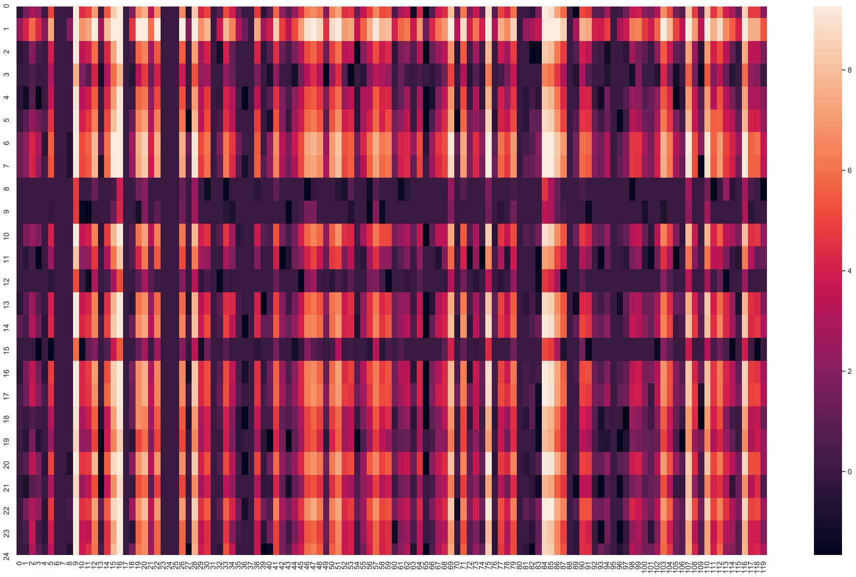

Figure 3.4: Predicted values, 2005.

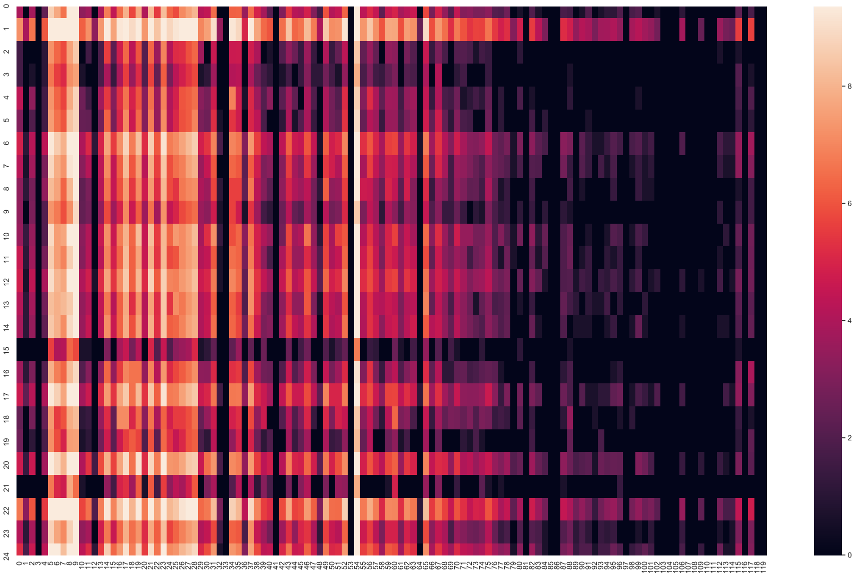

Figure 3.5: Actual values, 2005.

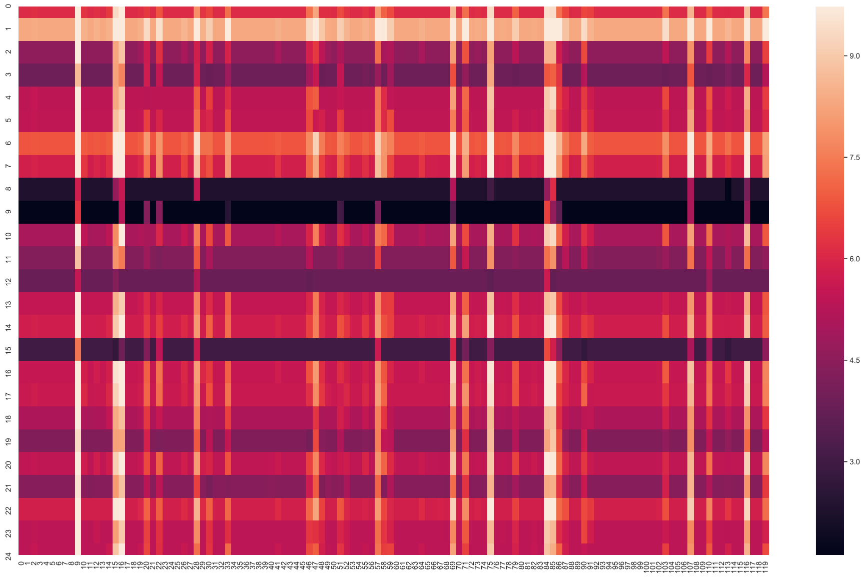

Figure 3.6: Predicted values, 2015.

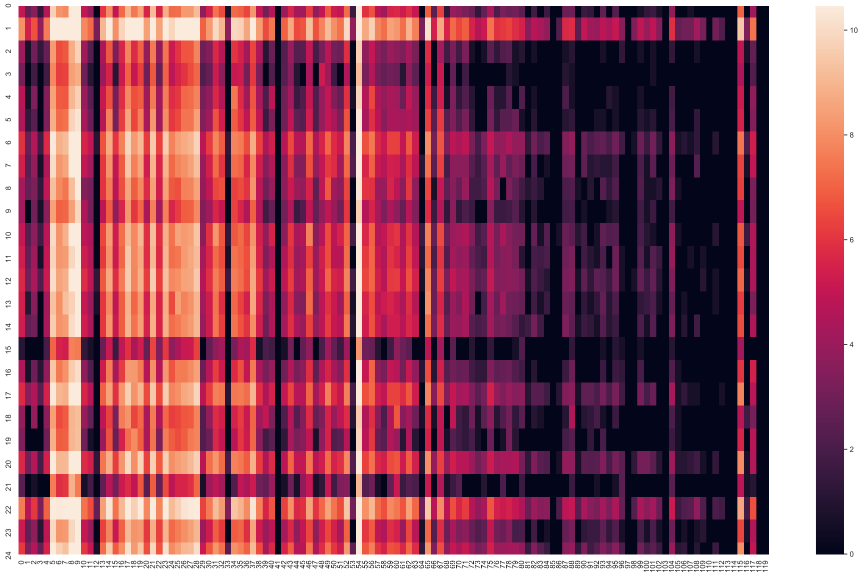

Figure 3.7: Actual values, 2015.

Unsurprisingly, the heatmaps representing 2005 values are much more similar than those for 2015. In 2005, a homogeneous increase in technology was predicted while the actual data suggest that certain terms and jobs experienced much less of an increase than others. In 2015, predictions of technological use were much too high. The model also failed to differentiate well between technological terms.

To explore the use of technology in each cluster of occupations, automation curves were constructed. Three different methods of visualization are presented. First, in figures 3.8 through 3.10, the raw co-occurrence counts are illustrated. Second, in figures 3.11 through 3.13, the log of raw counts are presented. Lastly, the percentage change in raw counts is represented in figures 3.14 through 3.16. For comparison, each set of figures presents the predicted values from 2005 to 2015, the actual values for 2005 to 2015, and the actual values from 1960 to 2015.

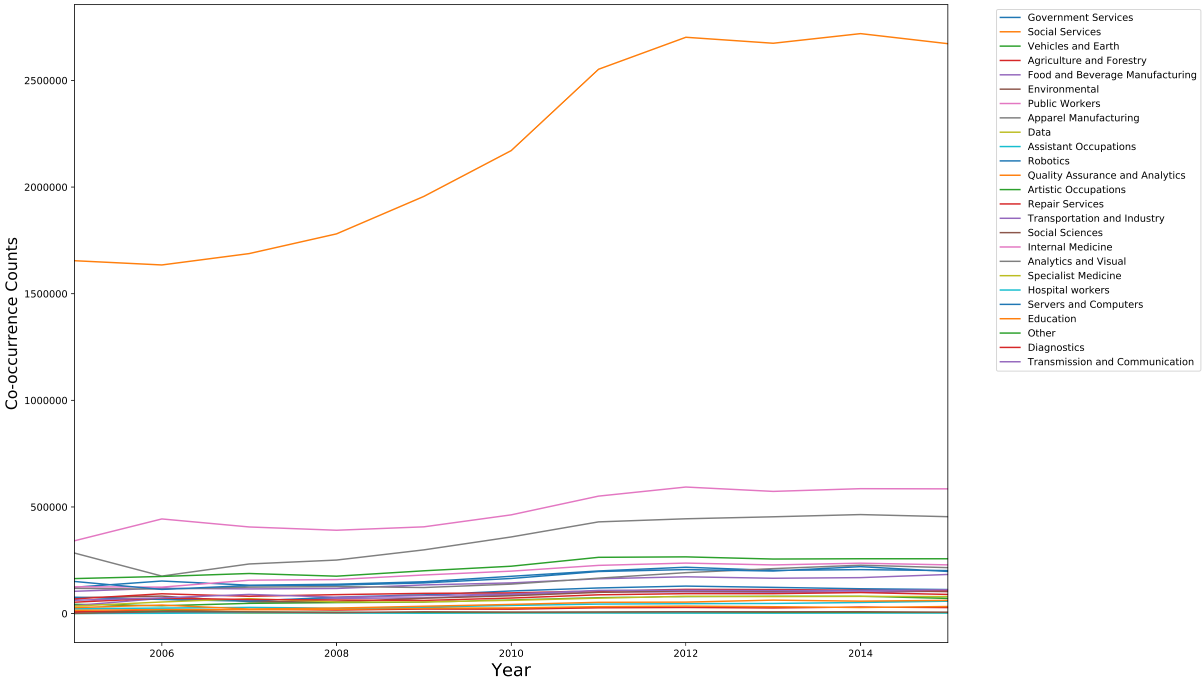

Figure 3.8: Predicted co-occurrence counts from 2005 to 2016, by occupation clusters.

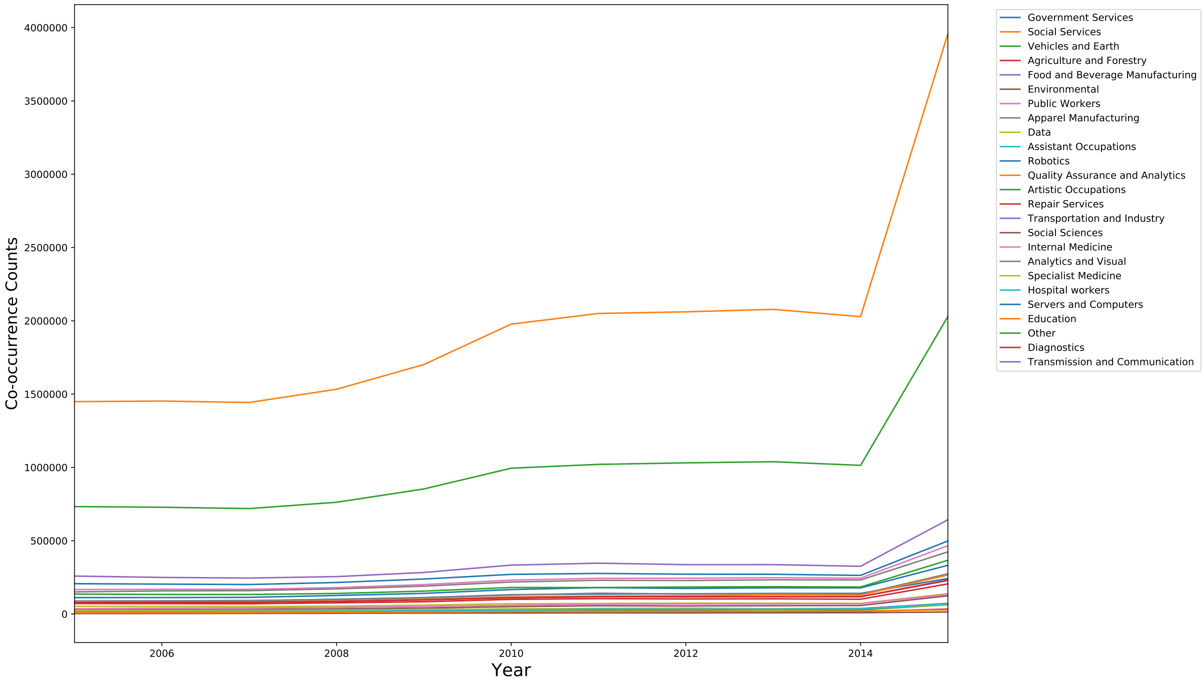

Figure 3.9: Actual co-occurrence counts from 2005 to 2016, by occupation clusters.

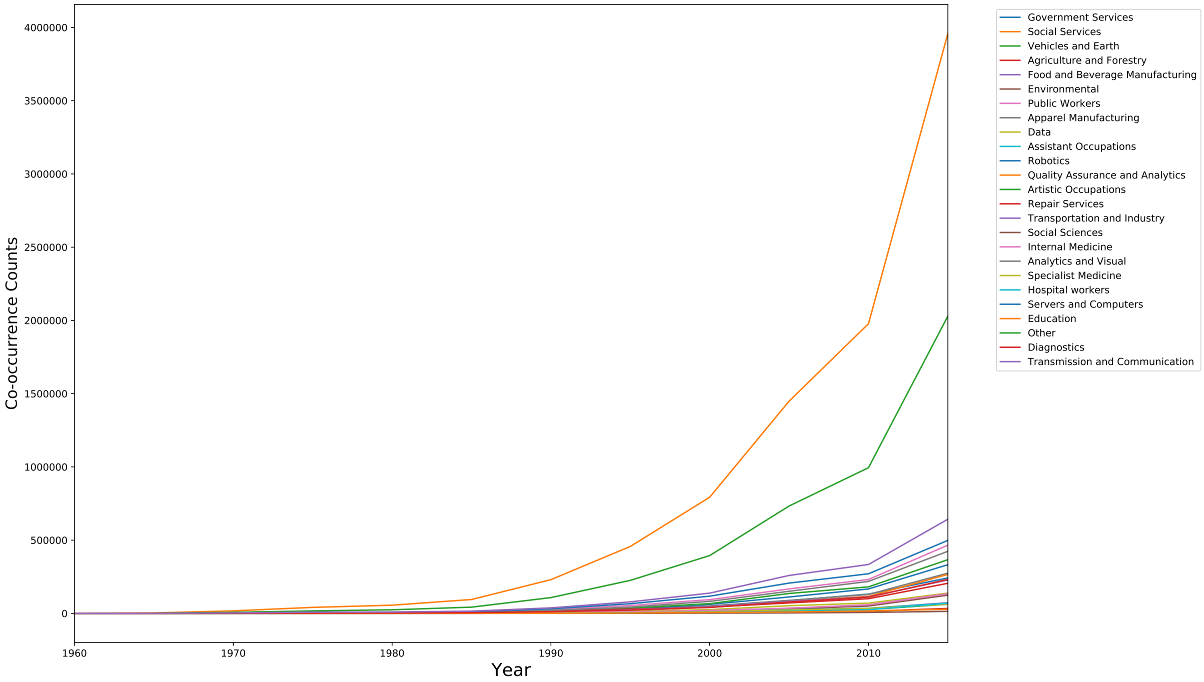

Figure 3.10: Actual co-occurrence counts from 1960 to 2016, by occupation clusters.

One immediate observation is that in for the predicted and actual values, most occupation clusters have less than 50,000 counts from 2005 to 2015, and the “Quality Assurance and Analytics” cluster has much higher counts than any other cluster in both. However, the “Vehicles and Earth” cluster is identified in the actual data as having higher counts, around 75,000 for the first years, but this is not predicted correctly. Further, the predicted values suggest that the “Public Workers” cluster and the “Analytics and Visual” cluster would have much higher counts than they actually do. Lastly, the model fails to predict the dramatic increase in counts seen around 2014. This is perhaps unsurprising, as such an increase seems to be unexpected given the trends seen in the preceding years.

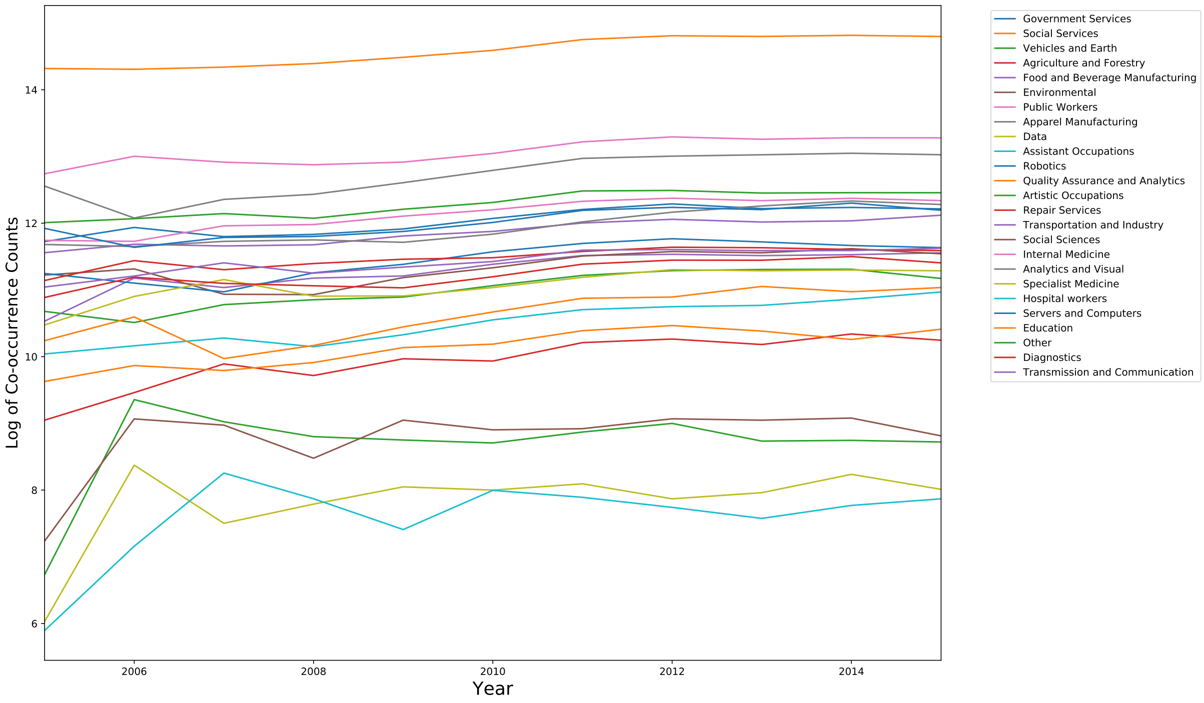

Figure 3.11: Log of predicted co-occurrence counts from 2005 to 2016, by occupation clusters.

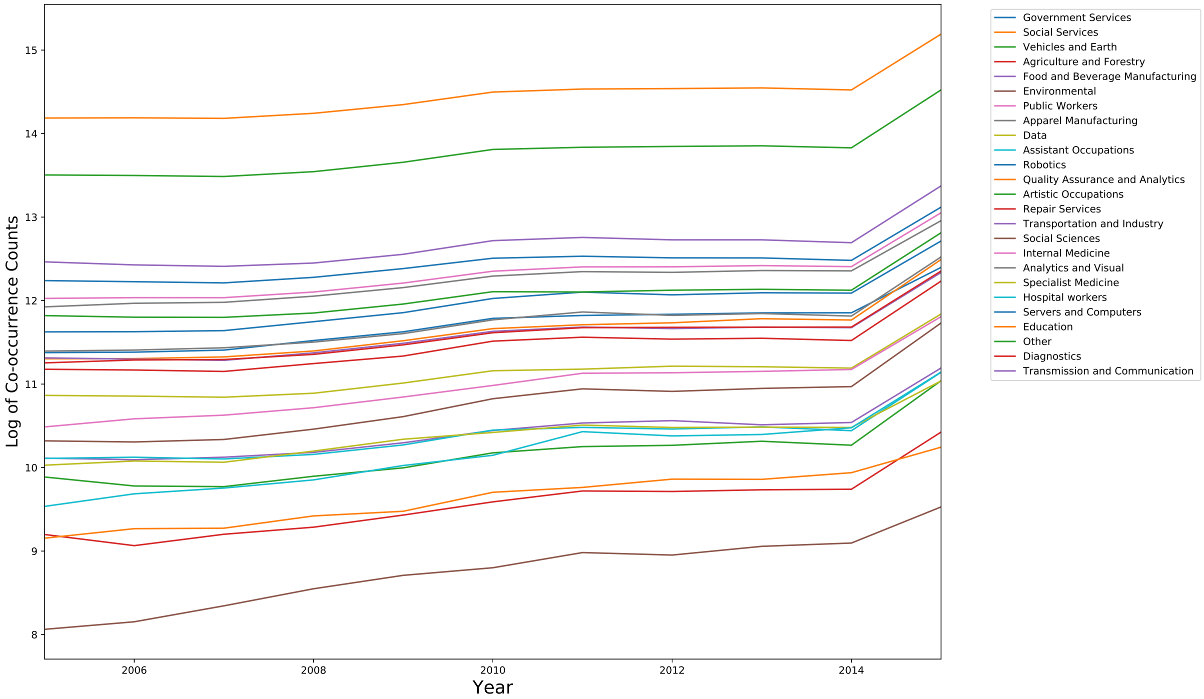

Figure 3.12: Log of actual co-occurrence counts from 2005 to 2016, by occupation clusters.

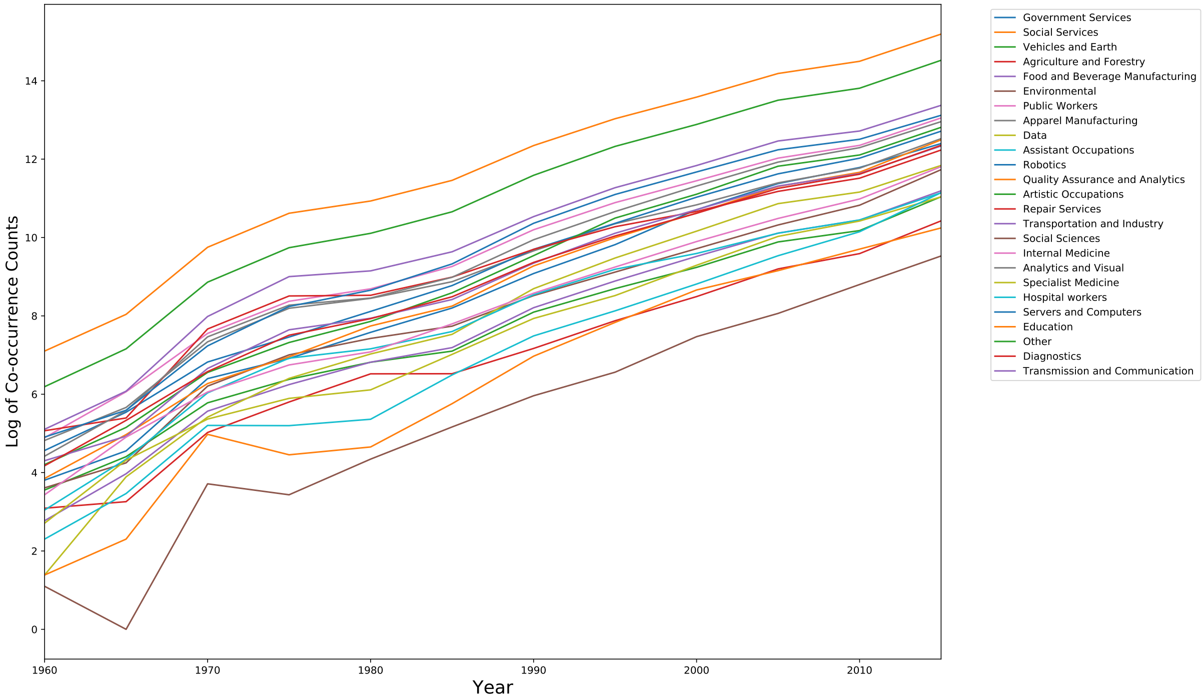

Figure 3.13: Log of actual co-occurrence counts from 1960 to 2016, by occupation clusters.

Visualizing the log of co-occurrence counts provides a more granular look at the occupation clusters that were seen as having less than 30,000 counts. These visualizations suggest that the predicted values are much more noisy than the actual data, especially in the earlier years of prediction. The general trend seems to be captured by the model, but the ordering of clusters is not consistent with that seen in the actual data.

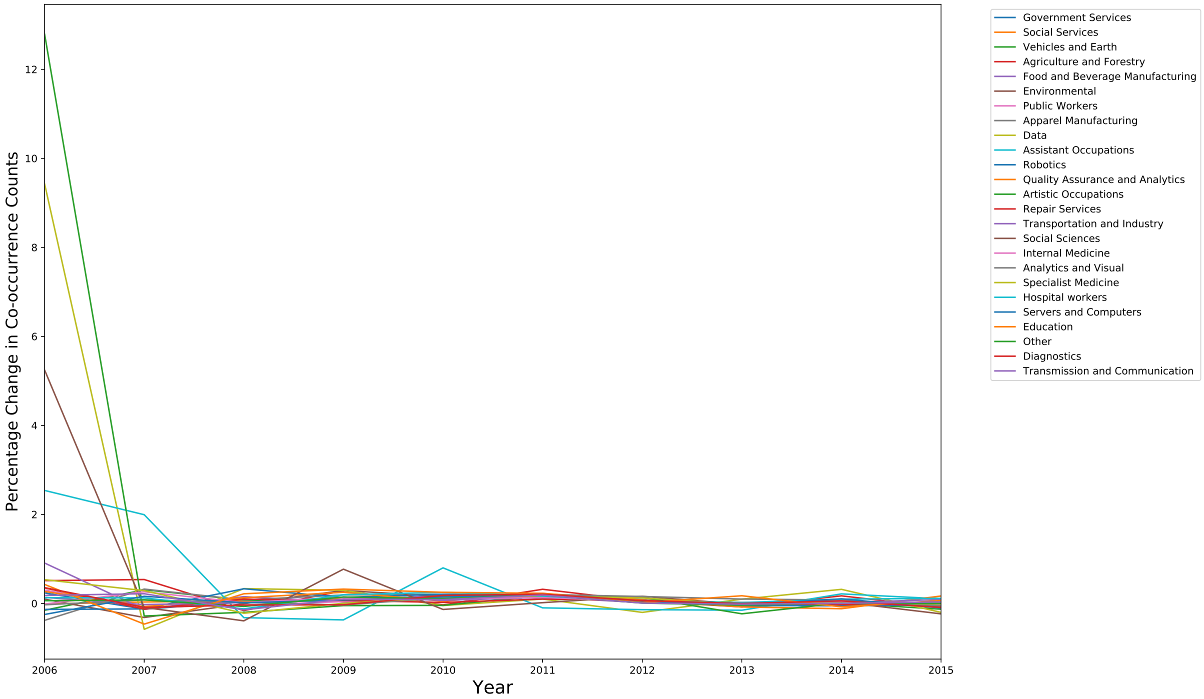

Figure 3.14: Percentage change in predicted co-occurrence counts, by occupation clusters.

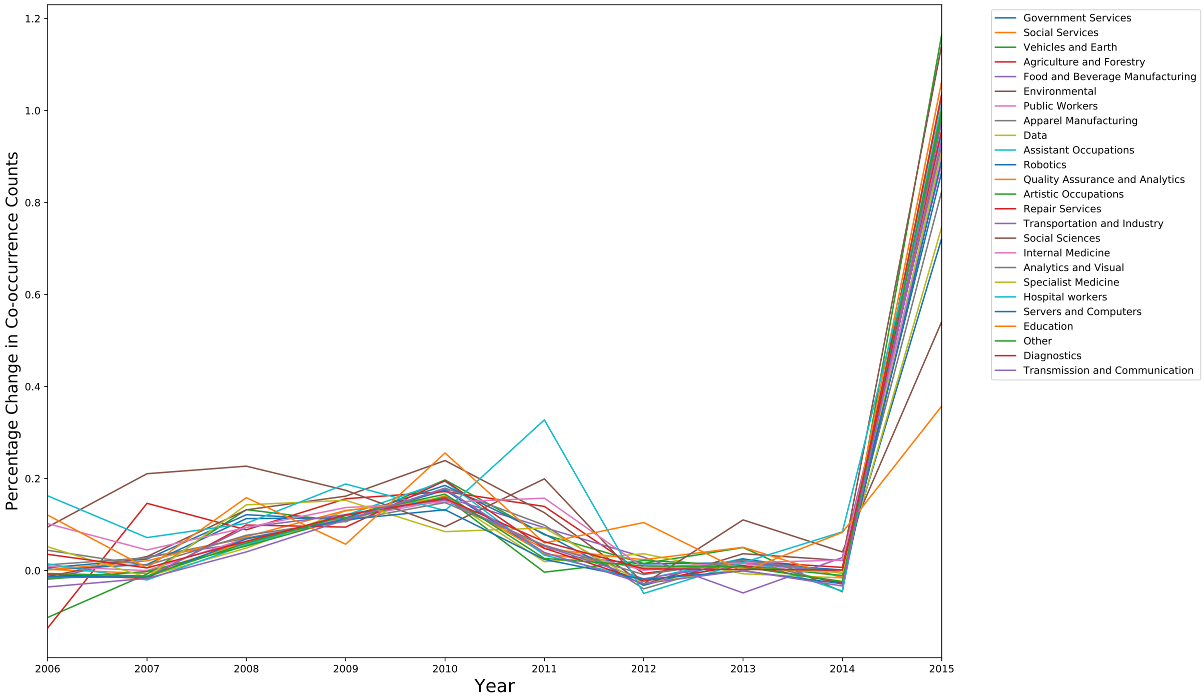

Figure 3.15: Percentage change in actual co-occurrence counts, by occupation clusters.

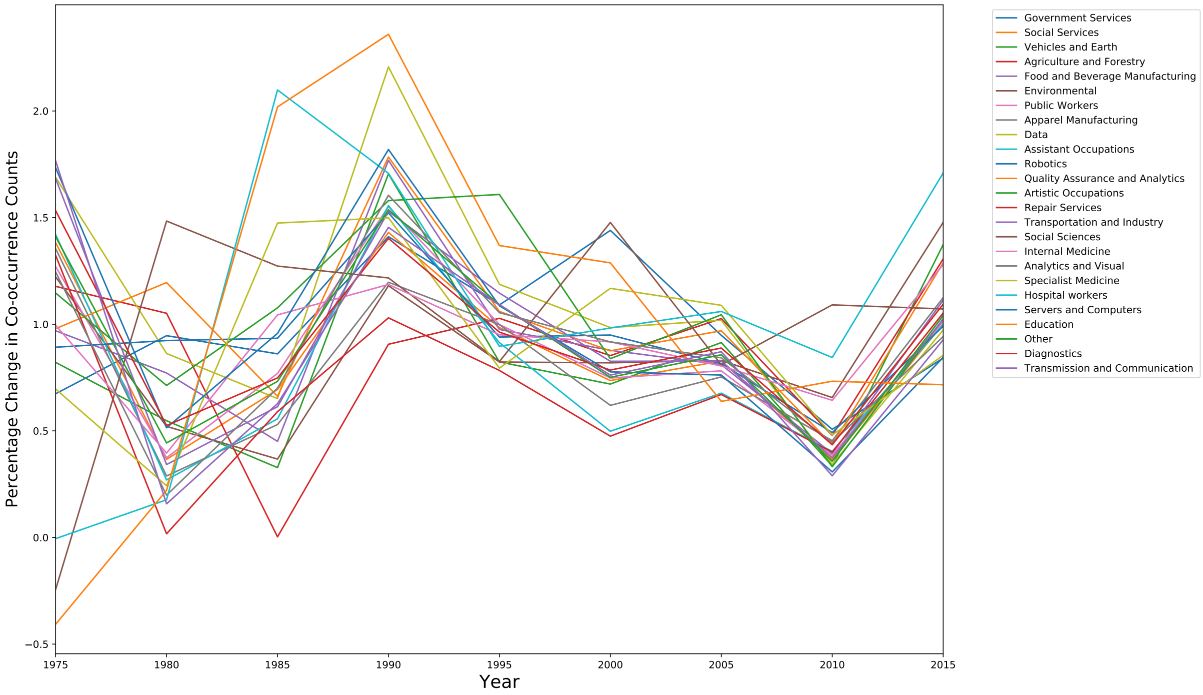

Figure 3.16: Percentage change in actual co-occurrence counts, by occupation clusters.

Examining the percentage change in values between years provides another look into the model's performance. Of note in these visualization is that the model tends to underpredict year-to-year changes. The predicted changes are all clustered around 0, while the actual changes are much more noisy and follow a general trend upward until 2010 before falling to 2014 and spiking after 2014. Since the model was predicting counts and not changes, its predictions are much similar between years than actual data would be.

Since many technological terms may appear in paper abstracts without any connection to any occupation terms that also appear in the abstract, there is a possibility of false positive introduced through the method of generating data. As a check against false positives, data are examined for only machine learning (ML) and artificial intelligence (AI) terms. Co-occurrence counts for these two terms are summed for each occupation cluster instead of summing co-occurrence counts for all terms, and the same visualization techniques are carried out. Figures 3.17 through 3.22 present the results. Figures 3.17 through 3.19 presents raw co-occurrence counts and figures 3.20 through 3.22 present the percentage change in co-occurrence counts.

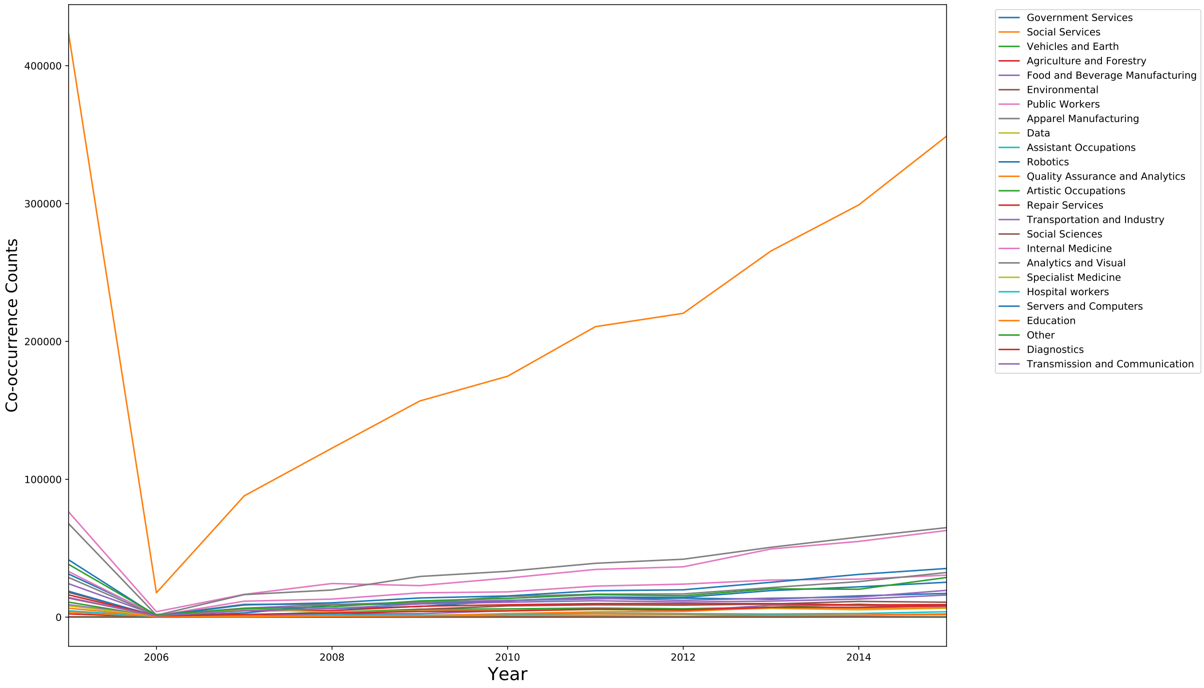

Figure 3.17: Predicted co-occurrence counts for AI and ML, by occupation clusters.

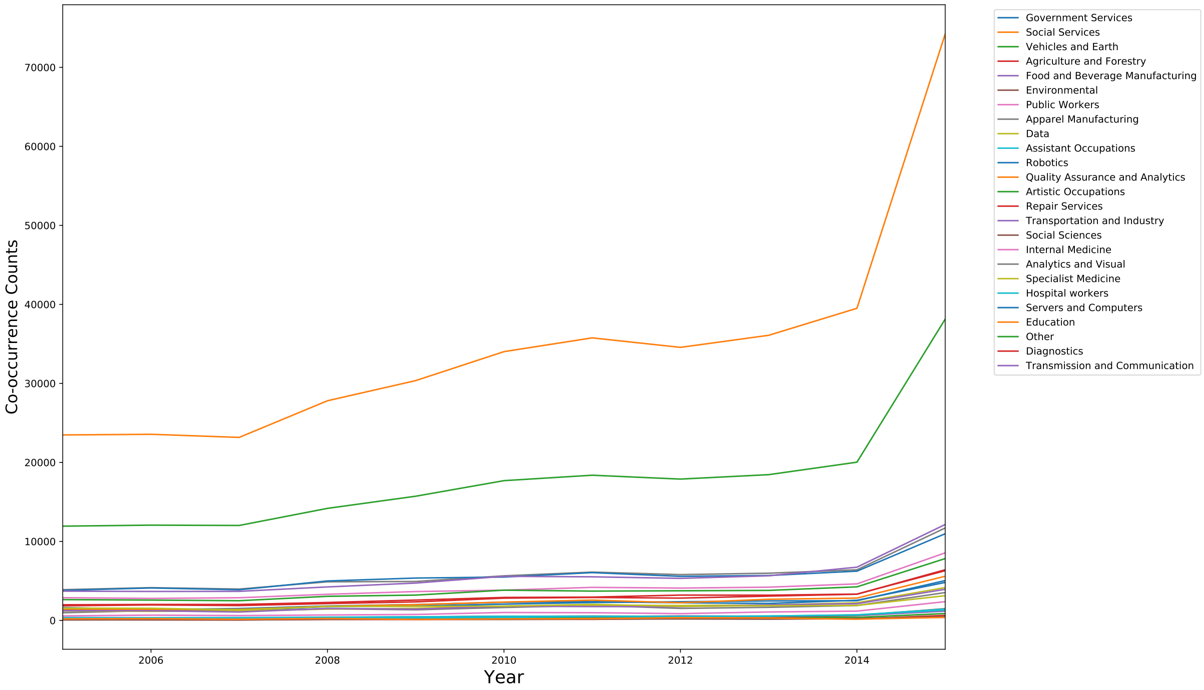

Figure 3.18: Actual co-occurrence counts for AI and ML, by occupation clusters.

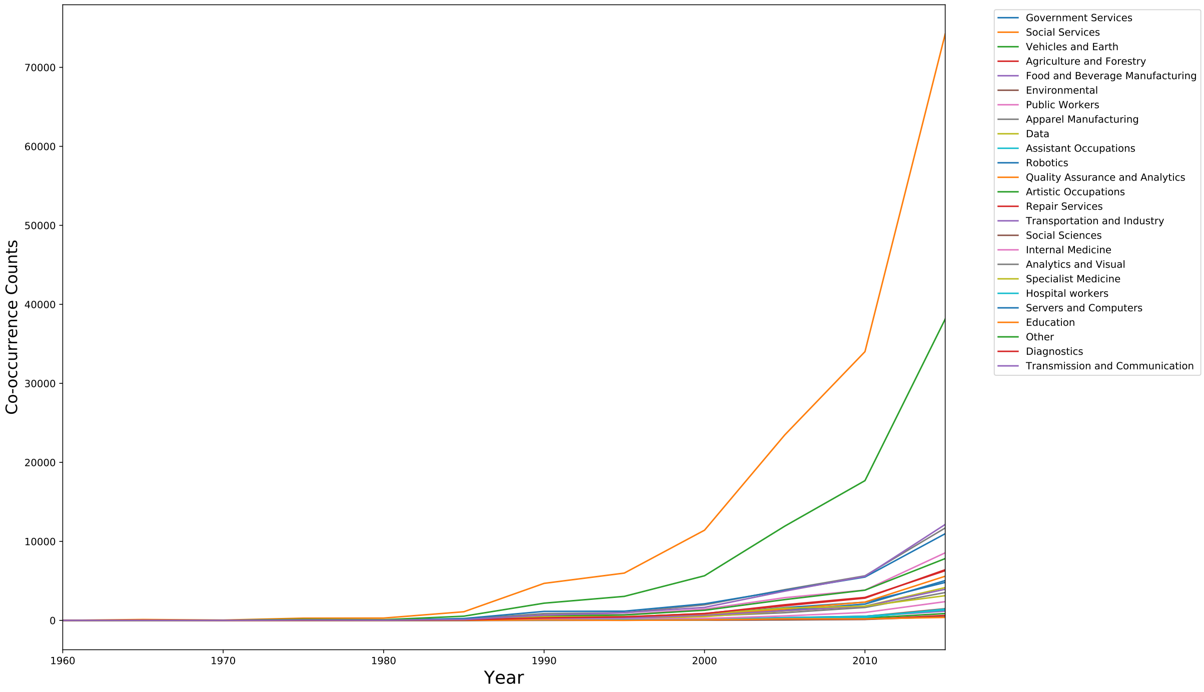

Figure 3.19: Actual co-occurrence counts for AI and ML, by occupation clusters.

Notably, the predictions for the majority of occupation clusters are similar to the actual values in that both are near zero for the majority of the time period. However, for the two clusters that have much greater co-occurrence counts, the model's performance is mixed. On one hand, it correctly predicts the higher counts for the cluster focused on Quality Assurance and Analytic occupations. On the other, it fails to identify the same trend seen in the "Vehicles and Earth" cluster of occupations, does not predict the dramatic uptick in 2015 seen in the actual data, and has original predictions for 2005 that are much to high.

To better represent the trends in the co-occurrence of machine learning and artificial intelligence terms and occupations, figures 3.20 through 3.22 present percentage changes in co-occurrence counts for machine learning and artificial intelligence. Notably, both the predicted and actual values are quite variable, and the changes in predicted values are not all that similar to the actual changes. This suggests that the model is better at predicting long-term trends than year-to-year changes, an unsurprising result due to the yearly and noisy nature of the data.

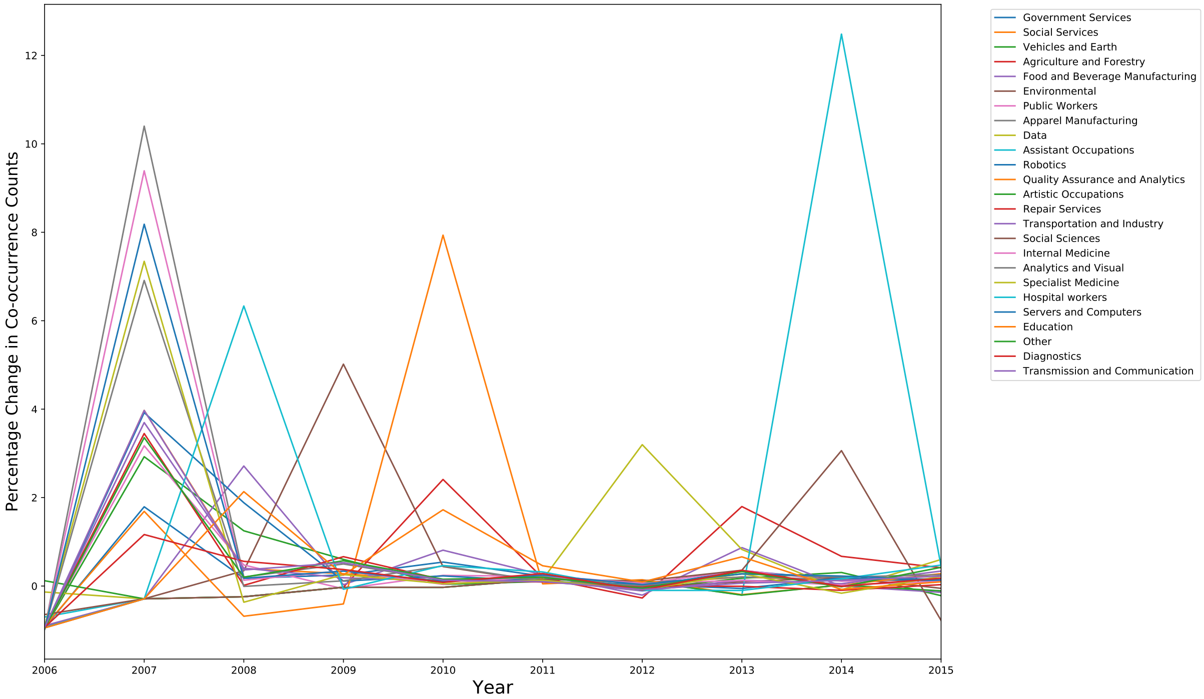

Figure 3.20: Percentage change in predicted ML and AI co-occurrence counts, by occupation clusters.

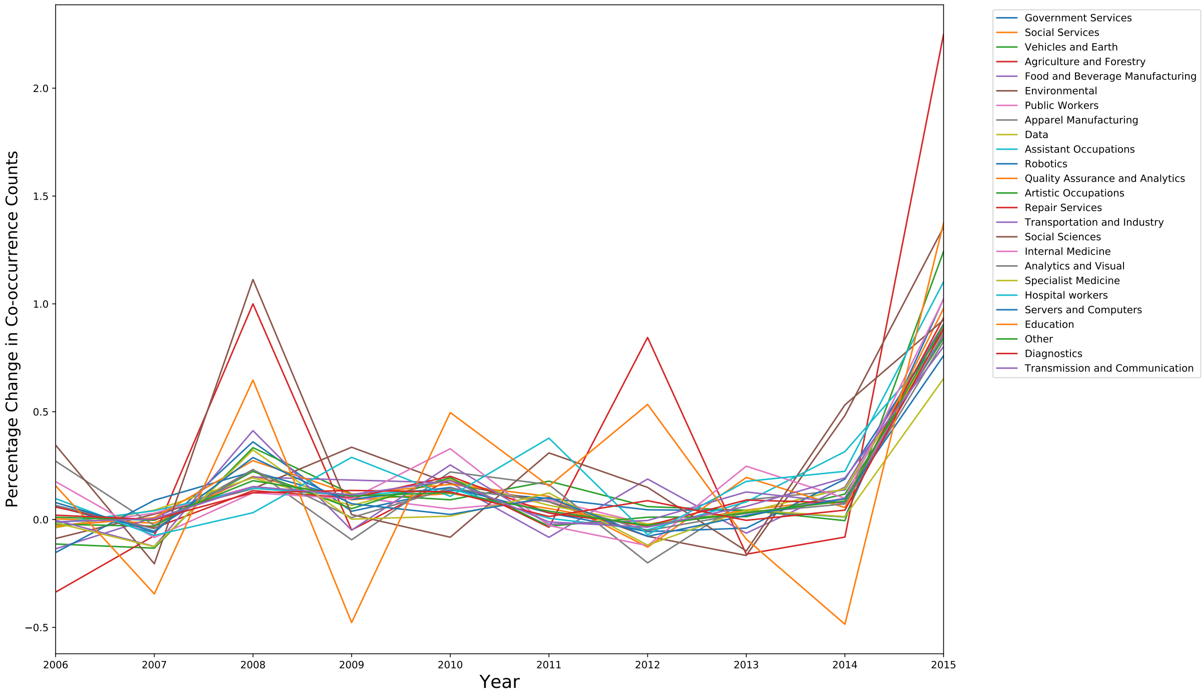

Figure 3.21: Percentage change in actual ML and AI co-occurrence counts, by occupation clusters.

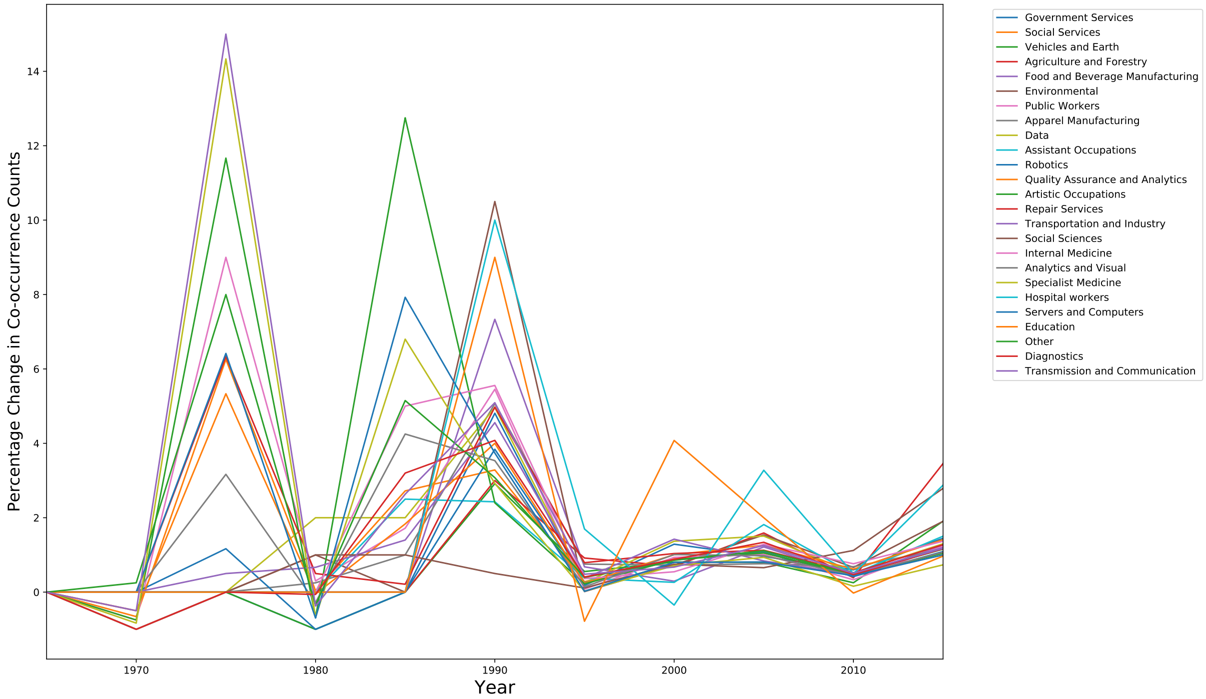

Figure 3.22: Percentage change in actual ML and AI co-occurrence counts, by occupation clusters.

Chapter 4

Discussion

The models explored above suggest that co-occurrences of references to technological terms and occupations in academic publications are relatively predictable from past co-occurrences. As seen in Figures 3.1 and 3.2, the MSE of predictive models drops to near zero when predicting counts using most of the years of data to train the model. Additionally, these models show higher accuracy than a baseline model using only one prior year to predict each year's values, especially in later years when more data is used to train.

However, visualizations suggest that the model becomes less reliable when used to predict multiple years, as shown in figures 3.4 through 3.7. Figures 3.4 and 3.5 suggest some discrepancies between actual and predicted values, seeming to project more homogeneity in increased usage across technology terms but generally being comparable to the real data. Figures 3.6 and 3.7 illustrate predictions that are ten years removed from the real data. Not only are the predicted values much higher than in reality, but much of the variation between technology terms and occupation clusters is lost. It is likely that this decrease in accuracy is caused by the inclusion of past predictions in the model's training data, which results in the accumulation of error.

Figures 3.8 through 3.22 also support the idea that while the model captures general trends in automation, it fails to capture more subtle trends such as year-to-year changes and the ordering of occupation clusters. As discussed in the limitations section, inaccurate clustering of occupations may play a role in this inaccuracy. Despite this, the curves follow a general pattern of increasing over time that matches the real world occurrences.

Section 4.1 Limitations

As a driving motivator of this paper was to develop a data-driven approach to predicting future automation, little manual work was done to cultivate the list of occupations and technological terms. While this is desirable in that it avoids bias, it also results in lists that may be sub-optimal. This is especially true for the list of occupations. Frequently, terms that would not be used in general language to refer to jobs were introduced by Word2Vec as similar words to the seed list. An additional problem arose with clustering the lists of terms. When clustering was performed based on words' embeddings, these words were grouped based on their alphabetical similarity and not necessarily because they refer to similar occupations or technologies. While this would not have a negative effect on the measures of model accuracy, it would be reflected in the heatmaps used to compare predicted and actual automation that are presented above.

Additional limitations arise through the method of using co-occurrence matrices to model automation. While it is likely that academic publications capture industry trends in the use of technology and thus are reflective of actually automation, this relationship is not an exact match. There are many other factors influencing industries' adaptation of technology, including cost, resistance among employees, and availability of said technology, that are not measurable through academic publications. Ideally, data capturing the extent of automation across industries could be included as training data in the above models, and academic publications could be used as predictors of actual automation. However, this data is not available for either the United States or other nations, making this approach impractical. Further, the method of counting co-occurrences, in which any appearance of an occupation and technological term in an abstract is considered a positive co-occurrence, may bias results if abstracts refer to the two terms without any actual connection between them.

Section 4.2 Future Work

As noted above, a major limitation of this work was the lack of data measuring the levels of automation across industries over the last decades. Should such data become available, a natural extension of the framework proposed would incorporate this data into the training of the predictive model and the evaluation of the model's predictions. This would allow for a more accurate assessment of the reliability of the predictions and of their use in forecasting future automation and labor market changes.

Chapter 5

Conclusion

This paper proposes a novel predictive framework trained on co-occurrence matrices generated from publicly available abstracts of academic publications. These models produce accurate predictions of future co-occurrence matrices, but fall short of accurate predictions reaching many years into the future. While the model is limited in its ability to make long term predictions, the similarity of predictions in the shorter term suggests that co-occurrences of occupation and technological terms in academic abstracts is a relatively predictable phenomenon that is fit for use as training data in predictive models. While this work is limited by the lack of automation data to link academic references to actual changes in industry practices and employment, should future data emerge the approach described above could prove to be a highly effective way of predicting future changes in employment patterns and the labor force.

Bibliography

Arntz, M., Gregory, T., and Zierahn, U. (2017). Revisiting the risk of automation. Economics Letters, 159, 157-160.

Autor, D. Journal of Economic Perspective. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of economic perspectives, 29(3), 3-30.

Autor, D., and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. American Economic Review, 103(5), 1553-97.

Bessen, J. E. (2016). How computer automation affects occupations: Technology, jobs, and skills. Boston Univ. school of law, law and economics research paper, (15-49).

Effendy, S., and Yap, R. H. (2017, April). Analysing trends in computer science research: A preliminary study using the microsoft academic graph. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 1245-1250).

Ford, M. (2015). The rise of the robots: Technology and the threat of mass unemployment. Oneworld publications.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., and Wang, D. (2019). Toward understanding the impact of artificial intelligence on labor. Proceedings of the National Academy of Sciences, 116(14), 6531-6539.

Frank, M. R., Wang, D., Cebrian, M., and Rahwan, I. (2019). The evolution of citation graphs in artificial intelligence research. Nature Machine Intelligence, 1(2), 79-85.

Frey, C. B., and Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. Technological forecasting and social change, 114, 254-280.

Hahn-Powell, G., Valenzuela-Escárcega, M. A., and Surdeanu, M. (2017, July). Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. In Proceedings of ACL 2017, System Demonstrations (pp. 103-108).

Herrmannova and Knotch (2016), "An Analysis of the Microsoft Academic Graph". D-Lib, 22(9/10). (http://www.dlib.org/dlib/september16/herrmannova/09herrmannova.html).

Luo, F., Valenzuela-Escárcega, M. A., Hahn-Powell, G., and Surdeanu, M. (2018, June). Scientific Discovery as Link Prediction in Influence and Citation Graphs. In Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12) (pp. 1-6).

McAfee, A., and Brynjolfsson, E. (2017). Machine, platform, crowd: Harnessing our digital future. WW Norton and Company.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B., and Wang, K. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 990-998.

Vermeulen, B., Kesselhut, J., Pyka, A., and Saviotti, P. P. (2018). The impact of automation on employment: just the usual structural change?. Sustainability, 10(5), 1661.