THE STEM REQUIREMENTS OF "NON-STEM" JOBS:
EVIDENCE FROM UK ONLINE VACANCY POSTINGS AND IMPLICATIONS FOR SKILLS & KNOWLEDGE SHORTAGES

INNA GRINIS*

ABSTRACT. Do employers in "non-STEM" occupations (e.g. *Graphic Designers, Economists*) seek to hire STEM (Science, Technology, Engineering, and Mathematics) graduates with a higher probability than non-STEM ones for knowledge and skills that they have acquired through their STEM education (e.g. *"Microsoft C#", "Systems Engineering"*) and not simply for their problem solving and analytical abilities? This is an important question in the UK where less than half of STEM graduates work in STEM occupations and where this apparent leakage from the "STEM pipeline" is often considered as a wastage of resources. To address it, this paper goes beyond the discrete divide of occupations into STEM vs. non-STEM and measures STEM requirements at the level of *jobs* by examining the universe of UK online vacancy postings between 2012 and 2016. We design and evaluate machine learning algorithms that classify thousands of keywords collected from job adverts and millions of vacancies into STEM and non-STEM. 35% of all STEM jobs belong to non-STEM occupations and 15% of all postings in non-STEM occupations are STEM. Moreover, STEM jobs are associated with higher wages within both STEM and non-STEM occupations, even after controlling for detailed occupations, education, experience requirements, employers, etc. Although our results indicate that the STEM pipeline breakdown may be less problematic than typically thought, we also find that many of the STEM requirements of "non-STEM" jobs could be acquired with STEM training that is less advanced than a full time STEM education. Hence, a more efficient way of satisfying the STEM demand in non-STEM occupations could be to teach more STEM in non-STEM disciplines. We develop a simple abstract framework to show how this education policy could help reduce STEM shortages in both STEM and non-STEM occupations.

Keywords: STEM Education, Skills Shortages, Machine Learning

Date: First version: June 2016. This version: January 2017.

*Department of Economics, CEP and SRC, London School of Economics. Email: I.Grinis@lse.ac.uk.

I am very grateful to *Burning Glass Technologies* (BGT), especially Hal Bonella, Julia Schreiber and Bledi Taska, for giving me access to the BGT data, and to John Van Reenen, Roland Grinis and participants of the 1st IZA Workshop on the Economics of Education, the 8th Oxford Education Research Symposium, the Systemic Risk Centre research meetings and the LSE Labour work-in-progress seminars for interesting and helpful discussions, comments, and suggestions. This work is part of my Ph.D. research, kindly funded by the Economic and Social Research Council. All errors are mine.

1. Introduction

> *"A whole range of STEM skills - from statistics to software development - have become essential for jobs that never would have been considered STEM positions. Yet, at least as our education system is currently structured, students often only acquire these skills within a STEM track."*
>
> Matthew Sigelman [47]

To what extent do recruiters in non-STEM occupations (e.g. *Graphic Designers, Artists*, *Economists*) require and value knowledge and skills that, within the UK education system, are typically acquired in STEM (Science, Technology, Engineering, and Mathematics) disciplines?

Addressing this question is important because in the UK less than half of STEM graduates work in STEM occupations.[1] This apparent leakage from the "STEM pipeline" is often considered as problematic since STEM education is more expensive and difficult to acquire than non-STEM one.[2] Hence, if recruiters in non-STEM occupations do not really require and value STEM knowledge and skills and simply like hiring STEM graduates for their "foundation competencies" (Bosworth et al. [9]), "logical approach to solving problems" (BIS [42]) or just because they believe that STEM graduates are more capable than their non-STEM fellows, the UK may be wasting a lot of money and efforts.

Another possibility, however, is that the discrete divide of occupations into STEM vs. non-STEM is imperfect and does not capture the changing nature of the UK economy, hit by trends like digitization, the arrival of Big Data, etc. which transform business operations and infiltrate STEM requirements throughout the economy and, in particular, outside positions that are typically considered as STEM.[3]

Indeed, "STEM occupations" are a relatively arbitrary construct. They are identified using judgment (Mason [41], BIS [42], BIS [8], Greenwood et al. [30], DIUS [17], Chevalier [14]), data-driven approaches (Bosworth et al. [9], Rothwell [45]), or a combination of both (UKCES [23]). Most studies recognize that "the issue of precisely where to draw the line between STEM and non-STEM never goes away" (Bosworth et al. [9]), that "neither Standard

[1]This finding is robust to different ways of defining STEM occupations and STEM disciplines, e.g. Chevalier [14] examines the LDLHE and finds that 36% of scientific graduates work in scientific occupations six months after graduation. The proportion is 46% three and a half years after graduation. Bosworth et al. [9] find that core STEM occupations employ only 40% of core STEM degree holders.

[2]Most STEM subjects fall into price categories A to C and therefore receive more funding from the Higher Education Funding Council For England (HEFCE) than the majority of non-STEM subjects which belong to price group D (HEFCE [18]).

[3]For example, see Brynjolfsson & McAfee [12] for a review of how Big Data is transforming management practices.

Occupational Classification (SOC) system codes nor Standard Industrial Classification (SIC) codes are particularly valuable to [classify STEM employment]" (BIS [42]), and that "STEM degree holders working in a non-STEM occupation may still be using their STEM skills" (Bosworth. et al. [9]).[4]

The only way to shed more light on this important issue is to go beyond *occupations* and measure STEM requirements at the level of *jobs*. We shall attempt this by examining the universe of UK online vacancy postings. Our data comes from *Burning Glass Technologies* (BGT), a labour market analytics company that collects, deduplicates and processes information on all UK online vacancies posted on employer websites, major job boards, government databases, etc. Where available, BGT collects job titles, occupation, industry and employer identifiers, education, experience and discipline requirements, wages, geographical locations ... and, most importantly, transforms the job description texts into sets of keywords, e.g:

"SAS - Writing - Data Collection - Econometrics - Project Design - Team Building - SQL - R"

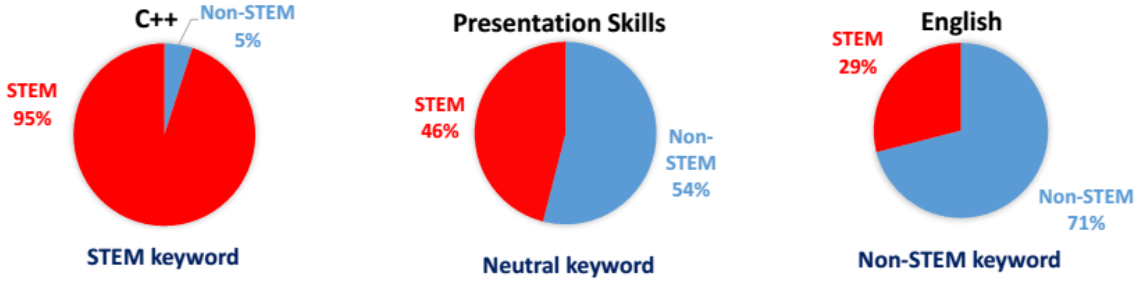
"Financial Analysis – Photography – Rehabilitation"

Our goal is to identify STEM jobs as those "involving activities that can only be satisfactorily carried out by individuals with STEM skills" (Bosworth et al. [9]). A straightforward approach would therefore consist in classifying as STEM those vacancy postings that explicitly require applicants to possess a STEM degree/qualification. However, unfortunately, only around 12% of all vacancy postings in our data contain any explicit discipline requirements. This happens because most UK recruiters prefer to simply describe the open position and the candidate that they are looking for directly, employing thousands of different keywords.

Hence, instead of relying on whether or not the posting contains an explicit STEM qualification requirement, we start by identifying "STEM keywords" - skills and knowledge that are

[4]Mason [41] applies judgment to the list of SOC occupations to identify those in which "the application of scientific, engineering and/or technological skills and knowledge is central to the job-holder's work". The list in Greenwood et al. [30] "was classified by a panel drawn from across the STE subjects and disciplines and convened by The Royal Academy of Engineering". BIS [42] rely on previous studies, their own judgment, preliminary discussions with key organisations and employer interviews to classify occupations into STEM Core, STEM related, STEM unrelated, and sectors into STEM Specialist, STEM Generalist and non-STEM. Bosworth et al. [9] use the Labour Force Survey to classify an occupation as STEM "if at least 15 per cent of its workforce is a STEM degree holder and the occupation as a whole employs at least 0.5 per cent of the STEM workforce." However, the problem with using the percentage of STEM degree holders as an indicator for whether or not an occupation is STEM, is that STEM graduates may be attracted to an occupation for reasons that are unrelated to employers' demand for STEM knowledge & skills, e.g. high wages. Recognizing this, UKCES [23] complement the proportion of STEM graduates in an occupation with a combined index for numeracy and problem solving skills use based on indicators from the UK Skills and Employment Survey. The "objective analysis" based on these two indicators outputs a list of 61 occupations. UKCES then refine this list using judgment. For instance, they remove health/medical occupations, teaching occupations and aircraft pilots as irrelevant, while including other occupations that were not identified as STEM in the objective analysis but "seem to be core STEM", e.g. technicians. Rothwell [45] uses O*NET Knowledge scales.

Figure 1.1 *Intuition behind "Context Mapping"*



Notes: Percentages of STEM vs. non-STEM discipline requirements with which a given keyword appears in the subsample of vacancy postings with explicit discipline requirements.

exclusively or much more likely to be taught in STEM disciplines (e.g. *"Systems Engineering"*), or job tasks, tools and technologies for which a STEM education is typically required (e.g. *"C++"*, *"Design Software"*) - using a method that we call "context mapping". The key idea in "context mapping", illustrated in Figure 1.1, is to classify keywords based on their "steminess" - the percentage of STEM discipline requirements with which a keyword appears in the subsample with explicit discipline requirements.

We then propose and evaluate several different ways of employing the steminess of all keywords in a given vacancy posting to classify it as STEM or non-STEM, as well as estimate the probability that its recruiter looks for a STEM graduate.[5] Our preferred classification method uses the steminess of keywords from both the vacancy description and the job title, and achieves an over 90% correct classification rate when tested on the subsample with explicit discipline requirements, i.e. classifies a job as STEM when the qualification requirement is a STEM discipline and as non-STEM when it is a non-STEM one.

Using this method, we classify all 33 million vacancy postings collected by BGT between January 2012 and July 2016. We find that around 35% of all STEM jobs belong to non-STEM occupations. Of course, nothing prevents STEM graduates working in non-STEM occupations to choose non-STEM jobs. However, if most of them happen to take up STEM employment opportunities, the fact that over half of STEM graduates work in non-STEM occupations may not be as problematic as often thought.

The list of non-STEM occupations with relatively high percentages of STEM jobs is very diverse and includes *Chartered architectural technologists* (85.42% in 2015), *Production managers and directors in construction* (78.56%), *Business, research and administrative professionals n.e.c.* (46.84%), *Product, clothing and related designers* (45.62%), and even *Artists* (23.46%).

[5]For simplicity, throughout this paper, we often use the expression "STEM graduate" to mean "STEM educated candidate" at any education level. Similarly, we use interchangeably the words "recruiter", "vacancy", "job", "posting".

Perhaps surprisingly for the literature, where financial occupations are typically considered as the main non-STEM group poaching STEM graduates, none of them is actually top of the list in terms of the percentage of jobs classified as STEM. For instance, among the seven occupations defined as financial in Chevalier [14], *Management consultants and business analysts* was the one with the highest percentage of STEM jobs in 2015: 25.33%, followed by *Financial and accounting technicians* with only 11.67%. The reason may be that, within the UK education system, the "numerical skills" for which financial occupations are thought to be seeking STEM graduates are actually also often transmitted to non-STEM graduates in, e.g., *Finance* or *Economics* degrees. Hence, although numerous jobs in financial occupations may end up being filled with STEM graduates, when posting their vacancy, few financial recruiters describe the job as one that could only be undertaken by someone with a STEM education.

As expected, most of the jobs within STEM occupations are identified as STEM (81% of all), while, in non-STEM occupations, STEM jobs remain a minority - around 15% of all. However, even these small percentages add up to a significant number of STEM employment opportunities outside STEM occupations and ignoring them leads to underestimating the overall demand for STEM knowledge and skills. For instance, in 2015, 2.66 million STEM vacancies were advertised online, while the number of jobs posted in STEM occupations was only 2.15 million. Hence, equating STEM jobs with STEM occupations would make us underestimate STEM demand by around half a million vacancies.

An important argument often put forward to defend the view that the breakdown of the STEM pipeline is problematic, is that STEM graduates receive a wage premium only if they stay in STEM occupations, i.e. "STEM skills are not particularly valued in non-STEM occupations" (Levy and Hopkins [38]). The evidence often mentioned is the DIUS report [17] which finds that "science graduates who work in science occupations earn a wage premium even allowing for other factors. [...] The remainder work in other occupations where they may well be using the analytical skills acquired during their education; however, they do not earn a higher wage in these occupations than equivalent people who studied other subjects".[6]

In reality, the DIUS report uses the LDLHE and the Labour Force surveys and therefore cannot distinguish between STEM and non-STEM jobs within non-STEM occupations. By contrast, our approach allows us to make this distinction. Although our results are not directly comparable because we examine the wage premium for STEM from the labour *demand* side and we do not claim causality and caution that our results could be biased because of

[6]A similar conclusion is reached by Chevalier [14] who also uses the LDLHE. Accounting for selection into both science degrees and science occupations, he finds that the returns to a science degree are small at 2% (and significant at 10% only) and are dominated by the returns to a scientific occupation at 18% (and highly significant). The findings in Greenwood et al. [30] are more nuanced. They analyse the Labour Force Survey between March 2004 and December 2010, and find that "degrees in STEM are valued by the labour market anyway but particularly so in STE occupations."

some omitted unobserved heterogeneity akin to the "ability bias" on the labour supply side, we find that STEM jobs are associated with higher wages both within STEM and non-STEM occupations. The premium remains significant and large even after controlling for a full set of four-digit occupations, education and experience requirements, counties, one/two digit industries, employers, etc. Moreover, conditional on a full set of four-digit occupation fixed effects, there is no statistically significant difference between the wage premium offered for STEM knowledge and skills in STEM occupations and the one offered in non-STEM ones.

Note that our results do not necessarily contradict but rather extend previous evidence because, within non-STEM occupations, nothing prevents STEM graduates to take up non-STEM jobs, for which non-STEM graduates are also perfectly qualified and no premium is offered. The distinction with previous studies is the finding that around 15% of recruiters in non-STEM occupations do require STEM knowledge and skills and offer to pay a premium when doing so.

Overall, our empirical results therefore suggest that the leakage from the STEM pipeline may be less problematic than previously thought, because a significant proportion of jobs in non-STEM occupations can only be satisfactorily fulfilled with people possessing a certain level of STEM knowledge and skills, which, within the UK education system, is typically acquired through a STEM education. Moreover, our findings suggest that STEM shortages may exist not only in STEM occupations but also in non-STEM ones.

Nonetheless, the STEM pipeline breakdown remains problematic for two main reasons. Firstly, as already mentioned, many STEM graduates working in non-STEM occupations could still be taking up non-STEM jobs. More importantly, there could be more efficient ways of satisfying the STEM demand in non-STEM occupations than training full-time STEM graduates.

In fact, an interesting feature distinguishes STEM jobs in STEM occupations from their counterparts in non-STEM ones: while 60% of all keywords in the median posting of a STEM job in a STEM occupation are STEM, this number is only 30% for a STEM job in a non-STEM occupation. This suggests that STEM recruiters in non-STEM occupations are in reality looking for a certain combination of STEM and non-STEM knowledge and skills that lies in between the STEM-dominated combination required in STEM occupations and the predominantly non-STEM one asked for in non-STEM jobs.

A recent report by General Assembly & BGT [3] calls this type of jobs "hybrid" since they "blend skills from disciplines which are typically found in disparate silos of higher education." They identify six "hybrid" job categories, e.g.: *Marketing Automation*, which "blends marketing with information technology", *Product Managers* who "draw from both business / marketing and computer programming", *UI/UX Designers* who "call on skill sets from design, programming and even psychology or anthropology."

They also note that "while the market increasingly demands these skill cocktails, higher education programs have been slower to package learning in such cross-disciplinary ways." Indeed, the reason why we identify these jobs as STEM occurs precisely because their recruiters are still looking to hire STEM graduates with a higher probability than non-STEM ones. This may happen because, within the UK education system, non-STEM graduates are typically unqualified for such "hybrid" positions: even if they possess the required non-STEM skills, they do not master the STEM ones, which may be more difficult and/or expensive for the employer to train and are therefore a prerequisite.

However, digging further into the STEM requirements of "non-STEM" jobs (STEM jobs in non-STEM occupations), we find many skills and knowledge that could certainly be acquired through appropriate training that is less advanced than a full-time STEM degree - e.g. learning how to code in, say, *"C++"* or *"Python"* does not necessitate a Bachelor in Computer Sciences. This agrees with the General Assembly & BGT report which also emphasizes that these new hybrid roles "are accessible with technical training less than a computer science degree."

Although increasing the number of people studying STEM disciplines is one of the most popular solutions proposed to reduce STEM shortages (e.g. Rothwell [44]), our findings suggest that a more efficient way of satisfying STEM demand within non-STEM occupations may be to teach more STEM in non-STEM disciplines in order to make non-STEM graduates qualified for a set of jobs within non-STEM occupations for which they only lack the STEM skills while already possessing the required non-STEM ones. In Section 5, we construct a simple abstract framework to illustrate how STEM shortages in STEM and non-STEM occupations are related and why this reform could help alleviate both.

This paper inscribes itself in the literature that employs online vacancies' data to investigate labour market dynamics and/or inform education provision policies. Although this type of data comes with important caveats that we discuss in more details below in Section 2, it is highly valuable to both academics and policymakers because of its many advantages over the more constrained and costly surveys which rely on random sampling and are typically less detailed. Reamer [43], for example, gives an interesting overview of how real-time labour market information could be used by different federal agencies and trade associations in the US to better align education programs with current labor market demand. He also discusses the pros and cons of such usage.

The BGT data itself has already been employed for a variety of internal and external research projects.

In the UK, BGT have partnered with the Institute for Public Policy Research (IPPR) to create an online skills calculator that "compares entry-level employer demand and the number of learners completing related programmes of study".[7]

In the US, the paper most related to our work is Rothwell [44]. He uses a subset of the BGT data for which the duration of the vacancy is known to show that STEM job openings take longer to fill than non-STEM positions at all education levels. However, STEM jobs in Rothwell's paper are still identified at the *occupation* level. In particular, he uses O*NET Knowledge scales, as explained in his other paper [45] that we discuss in some details in Section 4. He does not seek to use the keywords from the vacancy descriptions to classify the job postings as STEM or non-STEM directly. Instead, he defines the value of each BGT keyword, called "skill", as the average salary cited in the postings containing it. He finds that more valuable skills are advertised for longer and that STEM positions tend to demand more valuable skills.

Several academic papers employ US BGT data to investigate the "upskilling" phenomenon over the business cycle. Ballance et al. [6] find that an increased availability of workers during downturns leads employers to raise their education and experience requirements. However, as the authors show in their next paper [5], the upskilling that happened during the Great Recession has been reversed as the labour market improved from 2010-2014. By contrast, Hershbein and Kahn [32] argue that Ballance et al. [5] "overstate the degree of downskilling during the later recovery" and provide evidence that the Great Recession was a time of "cleansing" during which many firms restructured their production in a manner consistent with routine-biased technological change, therefore increasing skill requirements permanently.

Hershbein and Kahn [32] use the keywords part of the BGT data to define "computer" and "cognitive" skill requirements. They designate an ad as requesting computer skills if it contains the keyword "computer" or one of the keywords categorized as "software" by BGT themselves (822 keywords in the UK taxonomy). They consider as "cognitive" skills all BGT keywords that contain "research", "analysis", "decision" and "thinking", e.g.: *"Online Research", "Logit Analysis", "Clinical Decision Support"*, etc. In the UK BGT taxonomy, which contains 11,182 distinct keywords overall, this amounts to 280 keywords. Hence, de facto, Hershbein and Kahn [32] classify less than 10% of all keywords as either "computer" or "cognitive" skill requirements ((280+822)/11182). The problem is that the unclassified 90% contain many keywords, like *"Algebra", "Machine Learning", "Natural Language Processing", "Graph-Based Algorithms"*, etc. which actually correspond to cognitive skill requirements without containing the four specific words that Hershbein and Kahn [32] focus on, and may also require computer skills without being included in the BGT software category. Also, note that the latter actually includes not only standard software like *"Microsoft Excel"* or

[7]http://wheretheworkis.org/

"MATLAB", but also many keywords that are not "computer" skills per se, e.g.: *"Flickr", "LinkedIn"*, *"Microsoft Live Meeting"*. Although the authors argue that they "ensure that the presence of these keywords correlates with external measures of cognitive skill at the occupation level", "many of [their] analyses exploit firm-level information", and at this more disaggregated level, such an incomplete classification of the BGT taxonomy could have tangible consequences. Moreover, on UK data, their approach gives some surprising results even at the occupation level with, e.g. 65.06% of *Economists* postings requiring cognitive skills in 2015, but only 44.39% of *Mathematicians* doing so.

Deming and Kahn [16] take a similar approach but go a bit further. This time the goal is to relate variation in skill demands to firm performance and wage variation within occupations. Although the authors argue that "the primary contribution of [their] paper is to distill and analyze the key words and phrases coded from the open text of ads in the BG data", in reality, they "distill" less than 20% of the BGT taxonomy by selecting the keywords that contain around 30 commonly occurring words and phrases, regrouped into 8 categories corresponding to different types of skills, e.g: cognitive, social, character, writing, etc. They also define computer and software skills based on the pre-existing BGT software category and the words "computer" and "spreadsheets" (cf. Table 1 in [16]).

Neither Hershbein and Kahn [32], nor Deming and Kahn [16] show why the fact that they work with such incomplete classifications of the BGT taxonomy does not affect their results.

In this paper, we also do not manage to classify all BGT keywords into STEM and non-STEM. However, we classify 85.55% of them and the remaining unclassified keywords appear very rarely in the postings so that, on average, 99.99% of all keywords collected from a vacancy with at least one keyword are actually classified. We further implement out-of-sample tests which recreate the situation of having a certain proportion of unclassified keywords to show that the number of misclassifications introduced by not being able to classify the remaining less than 15% is very small. Finally, we process the job titles into sets of keywords and add them to the BGT taxonomy, so that our eventual classification of jobs into STEM and non-STEM is based on 29,831 distinct keywords with 99.82% of all vacancies in our data possessing at least one classified keyword and the median number of classified keywords per vacancy with at least one being seven.

This paper also contributes to the emerging literature that develops and applies Machine Learning (ML) and Natural Language Processing (NLP) techniques to problems in Economics.[8] ML consists of "flexible, automatic approaches [...] used to detect patterns within the

[8]For instance, in labour economics, a recent paper by Frey and Osborne [25] also employs ML to examine the susceptibility of jobs to computerisation in the US. The authors hand-label 70 out of 702 US occupations as either automatable or not, then employ this sample to train a Gaussian process classifier and estimate the probability of computerisation for all 702 occupations as a function of nine O*NET variables that reflect bottlenecks to computerisation (e.g. finger dexterity, originality...). Their findings indicate that about 47 percent of total US employment is at high risk of computerisation (probability above 0.7).

data, with a primary focus on making predictions on future data" (Chiu [15]). It is becoming an indispensable toolkit for economists working with Big Data where standard approaches, like simply classifying a selected number of keywords, are not satisfactory and what is required from the researcher is to design, train and test algorithms that can automatically perform classification tasks on huge quantities of data.

The rest of the paper is structured as follows. We start by introducing the UK BGT data in Section 2 and explaining how STEM keywords and jobs are identified in Section 3. We then study the characteristics of STEM jobs in the UK in Section 4 and analyse the education policy implications of our empirical findings in Section 5.

2. Data

Nowadays, when wanting to hire someone, employers usually go online and post a job advert containing information about the vacancy they want to fill and the candidate they are looking for.

Burning Glass Technologies (BGT), a US labour market analytics company, has been collecting and processing information on all online job postings in the UK since 2012. Currently, they "spider" (visit) approximately 5,000 websites including major job boards (e.g. Career Builder, Universal Job Match), government job databases, direct sites of employers of all sizes and industries, as well as websites of agencies specialised in recruitment (e.g. Michael Page, Reed England).

BGT robots go online on a daily basis. However, the same vacancy ad spidered several times on the same or different platforms within a period of two months is removed as a duplicate. BGT regularly upgrades its infrastructure to enhance coding rules and expand posting sources, in which case it re-parses the entire database to ensure consistency and comparability of postings over time. The sample used in this paper runs from January 2012 to and including July 2016.[9]

Where available, BGT collects the job title, detailed information on occupation and industry identifiers, the employer, the geographic location, education, experience, and discipline requirements, wages, pay frequencies, salary types, and keywords from the job description texts. However, since few recruiters explicitly specify all this information in their vacancy postings and BGT does not impute any missing fields, the data contains many missing values.

Table 1 presents some summary statistics about the numbers of vacancies and the percentages of non-missing values in each year. Overall, our sample contains over 33 million of postings. Only 17.5% and 12.3% of them have minimum education and experience requirements respectively (the percentages are even lower for maximum requirements). The main reason is that employers often believe such information to be transparent from other

[9]The sample was received in September 2016, after the August 2016 update.

Table 1: Descriptive statistics, Jan. 2012 - Jul. 2016 BGT sample

	2012	2013	2014	2015	2016	Total
Panel A: Main Table						
Number of postings	5939705	7041917	6240340	8173962	5667039	33062963
% with Job Title	100	100	99.99	100	100	100
% with Occupation	99.73	99.54	99.44	99.51	99.48	99.54
% with County	95.55	88.88	80.04	77.8	79.66	84.09
% with Industry	47.08	45.78	46.96	45.37	45.06	46.01
% with Employer	24.86	29.73	30.93	31.85	32.2	30.03
% with Education (min)	16.24	18.28	19.02	17.27	16.85	17.56
% with Experience (min)	11.22	12.22	12.86	12.74	12.34	12.31
% with Salary	63.01	60.05	59.62	60.29	63.04	61.07
Hourly Salary (conditional on posting):						
Min	1.88	1.88	1.88	1.88	1.88	1.88
Max	72.12	72.12	72.12	72.12	72.12	72.12
Mean	15.58	16.10	16.50	17.17	17.21	16.54
Panel B: Keywords from Job Postings						
% with >= 1 Keyword	92.01	89.71	89.94	89.93	89.11	90.12
No. of unique keywords	9064	9496	9795	9995	9477	11182
Number of Keywords per Vacancy (conditional on posting at least one):						
Median	4	4	5	5	5	5
Mean	6.12	6.11	6.29	6.23	6.17	6.19
Max	226	211	115	111	167	226
Number of Vacancies per Keyword:						
Median	59	67	56	71	55	173
as % of all postings	0.001	0.001	0.001	0.001	0.001	0.001
Mean	3689.97	4067.25	3605.1	4580.86	3285.9	16482.43
as % of all postings	0.06	0.06	0.06	0.06	0.06	0.05
Most popular Keyword			*"Communication Skills"*			
% of postings	20.59	21.97	24.04	23.25	22.38	22.5
Panel C: Discipline Requirements						
% with >= 1 CIP major	11.43	12.04	13.14	11.75	11.74	12.01
			of which ...			
% with > 1 CIP major	30.72	29.54	30.22	31.04	31.53	30.58
% with >= 1 Keyword	98.87	98.68	98.83	98.76	98.53	98.74
No. unique CIP majors	394	402	403	403	398	425
No. of unique Keywords	8523	8831	8998	9026	8684	9566
as % of all Keywords	94.03	93.00	91.86	90.31	91.63	85.55
Number of Keywords per Vacancy in this subsample (cond. on >= 1):						
Median	8	7	8	8	8	8
Mean	9.47	9.04	9.17	9.13	9.21	9.19
Number of Vacancies per Keyword in this subsample:						
Median	27.50	29.00	26.00	28.00	25.00	67.00
Mean	871.38	963.66	920.79	1071.34	802.01	3767.79
% with non-mixed disciplines	90.72	91.00	90.60	90.46	90.34	90.63
Correlation with all postings:						
Keywords (No. times posted)	0.93	0.94	0.95	0.94	0.94	0.94
Occupations (4-digit SOC, %s)	0.82	0.81	0.80	0.80	0.81	0.81
County (%s)	0.99	1.00	1.00	0.99	0.99	0.99

Notes: Occupation (4-digit UK SOC), Industry (SIC at division or section levels), Education and experience requirements in years, Hourly salary (average of min and max if different). CIP stands for Classification of Instructional Programmes. % with non-mixed disciplines gives the % of vacancies for which all disciplines posted are either all STEM or all non-STEM.

characteristics of their job advert. For instance, the recruiter posting an *"Aerospace Engineer"* vacancy without an education requirement would not expect to receive applications from people with GCSE as the highest qualification. It should also be clear to the job seeker that the experience requirement of the vacancy whose title reads *"Vice President"* is different from the one with a title containing *"Analyst"*.

There are several other important caveats to bear in mind when working with online postings data. Firstly, some misclassifications are unavoidable when collecting data on such a grand scale. Moreover, not all vacancies are posted online, not all vacancies transform into real jobs, and sometimes a recruiter might post one vacancy but in reality seek to hire several people.

Despite all these shortcomings, occupational and geographic distributions in the BGT data exhibit high correlations with the occupational and geographic distributions of official UK employment data (the Annual Survey of Hours and Earnings (ASHE) from the Office for National Statistics (ONS)). Tables 17 and 18 in the Appendix present the results of a comparison analysis conducted by BGT for the 2014 sample. The correlation of distributions across major occupational groups is 0.94 (Table 17). However, as with US data, the UK data also exhibits an over-representation of positions typically requiring higher education (professional and associate professional occupations), and an under-representation of those requiring lower levels of education.[10]

In terms of geographic distributions, the correlations are also very high: 0.94 for professional occupations and 0.84 for elementary occupations (cf. Table 18). However, London postings are over-represented in the BGT sample.

Unfortunately, it is not possible to compare BGT data directly to the vacancies data from the ONS Labour Market Statistical bulletins because ONS uses three-month rolling averages (January-March, February-April, March-May, etc.), whereas BGT has a two-months deduplication window. Hence, a given posting in the BGT sample could appear more than once in ONS records. This may explain why, for instance, for 2014, ONS has 7.9 million vacancies, whereas BGT data contains only 6.24 million postings.

Moreover, it is important to remember that while BGT data contains the *universe* of online vacancies, both the ASHE and the Labour Market Statistical bulletins are based on *surveys* of households or businesses. For instance, the ASHE is based on a 1% sample of employee jobs, drawn from HM Revenue and Customs Pay As You Earn (PAYE) records. And as the ONS cautions itself, "results from sample surveys are always estimates, not precise figures."

2.1. **Keywords from Job Postings.** What makes BGT data stand out from more traditional sources of labour market information is the fact that it also contains keywords and

[10]For the US, Carnevale et al. [13] estimate that 80 to 90% of openings requiring at least a college degree are posted online, whereas the numbers for those requiring some college (or an Associate's degree) and those only requiring high school are 30-40% and 40-60% respectively.

phrases from the vacancy description texts. Concretely, in the data, the vacancy description text appears as a set of keywords taken out of context, e.g.:[11]

"Adobe After effects - E-Learning - Multi-Tasking - Audio Editing"

These keywords are collected using "a continuously expanding taxonomy" (Carnevale et al. [13]). We can think of this taxonomy as the "language" that recruiters employ to describe the job and the candidate they are looking for. It includes:

- Skills: *"Organisational Skills", "Time Management", "Communication Skills"...*
- Job tasks: *"Advertising Design", "Invoice Preparation", "Lesson Planning"...*
- Work styles: *"Detail-oriented", "Creativity", "Initiative"...*
- Software: *"Microsoft Office", "AJAX", "Adobe Acrobat"...*
- Knowledge: *"Civil Engineering", "Accountancy"...*
- Other: *"Her Majesty's Treasury", "FOREX", ...*

Any keyword in the job posting that has a match in the BGT taxonomy gets picked up. The order and number of times the keywords appear in the original job posting are ignored.[12] The taxonomy expands as BGT robots discover new keywords in job ads. Once new keywords are added to the taxonomy, all previous postings are re-examined to ensure consistency and comparability over time.

Currently, the taxonomy contains 11,182 distinct keywords, and 90% of all postings have at least one keyword (Panel B, Table 1). However, conditional on having at least one, the median number of keywords per vacancy is only 4-5. More importantly, in a given year, the median keyword appears in less than 0.001% of all postings. In fact, even the most popular keyword - *"Communication Skills",* appears in less than a quarter of all postings.

2.2. **Explicit Discipline requirements.** Only around 12% of all job adverts contain specific discipline requirements (Panel C, Table 1), e.g.: *"Chemistry", "Economics".*

The fact that most recruiters prefer to express their skills & knowledge requirements directly, by simply describing the open position and the candidate that they are looking for, is an important reason for attempting to identify STEM jobs from the vacancy description *keywords,* and not by relying on whether or not the posting contains an explicit STEM qualification/degree requirement.

However, since our goal is precisely to identify STEM jobs as those whose recruiters would most likely seek to hire STEM graduates, this sample with explicit discipline requirements

[11]BGT refers to them as "skills". However, because they also contain many expressions which strictly speaking are not "skills", we prefer to refer to both single word (e.g. *"Research"*) and multiple word phrases (*"Academic Programme Management"*) as simply "keywords". In practice, we removed the white space between the words in multiple word phrases to avoid treating, for instance, "Lotus Notes" and "LotusNotes" as distinct "keywords".

[12]Hence, the vacancy representation in our data is closer to what in the information retrieval literature is called a "boolean retrieval" rather than a "bag of words" model, although what is collected are specific keywords and phrases instead of all tokens (cf. Manning et al. [39]).

constitutes an important first step in our analysis. Within it, STEM jobs are already identified because we can directly observe whether the discipline posted is STEM or non-STEM.

Merging together observations for Jan. 2012 - Jul. 2016, we obtain almost 4 million vacancies with explicit discipline requirements. The 425 distinct disciplines posted in these 4 million vacancies correspond to majors from the *Classification of Instructional Programs* (CIP) - a taxonomic coding scheme of over 2,000 instructional programs, developed by the US Department of Education. The CIP has two-digit, four-digit, and six-digit series, and most of the programs are offered at the post-secondary level.[13]

We define STEM disciplines as the majors included in the CIP two-digit series corresponding to: Biological & Biomedical, Physical, and Computer **S**ciences, **T**echnology, **E**ngineering, and **M**athematics & Statistics. Table 16 in the Appendix provides the full list of disciplines contained within each group and that appear in our sample. All remaining disciplines in our data belong to different two-digit series and are therefore classified as non-STEM. Note that there is disagreement in the literature about whether Medical programs, Agricultural sciences, Environmental sciences and Architecture should be classified as STEM or not. In this paper, we decided to take the STEM acronym literally and therefore exclude these disciplines. However, future research could certainly explore alternative classifications.[14]

Around 30% of postings specify more than one CIP major. For such postings, we re-weight each major by the number of majors specified so that the overall discipline requirement sums to one.[15]

Figure 2.1 shows the resulting distribution of discipline requirements: 50.61% of CIP majors specified belong to non-STEM fields, while the rest are spread throughout the STEM domains, with 25.83% belonging to Engineering.

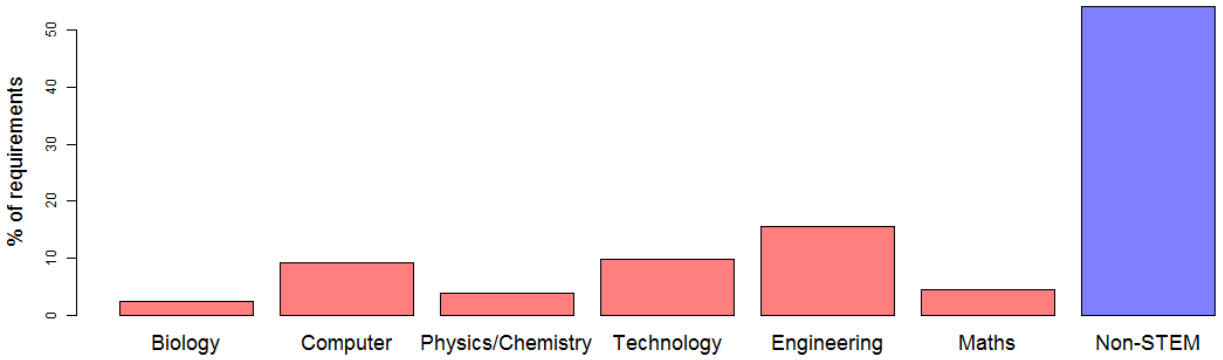
Only 9.27% of postings have mixed discipline requirements, i.e. specify CIP majors belonging to both STEM and non-STEM domains. 44.55% of vacancies have purely STEM discipline requirements and 46.17% have purely Non-STEM ones.

[13]A small proportion of the CIP corresponds to residency (dental, medical, podiatric, and veterinary specialties) and personal improvement and leisure programs; and instructional programs that lead to diplomas and certificates at the secondary level only. The latest 2010 edition of the CIP is available at: https://nces.ed.gov/ipeds/cipcode/Default.aspx?y=55. Note that the fact that the discipline is posted explicitly does not imply that the recruiter also specifies the minimum education *level* required. Indeed, in the sample with explicit discipline requirements, 38% of education level requirements are missing. 57.89% of those posted correspond to a minimum NQF level 6 or above (i.e. at least a Bachelor's degree).

[14]Similarly, although the US Department of Homeland Security (DHS) provides a list of CIP programs that it classifies as STEM, we decided not to use it because it has been created "for purposes of [a] STEM optional practical training extension" and contains a wide range of STEM-related disciplines in addition to the core ones, e.g. *"Educational Evaluation and Research"*. Moreover it is not directly comparable to the Joint Academic Coding System used in UK studies to classify disciplines as STEM or non-STEM.

[15]This ensures that we do not count such vacancies as many times as the number of disciplines that they specify instead of one, and also makes intuitive sense since a vacancy with two distinct discipline requirements is probably looking for a combination of knowledge and skills from both of them.

Figure 2.1 *Distribution of discipline requirements*



Notes: 3971988 vacancies with explicit discipline requirements collected between Jan. 2012 and July 2016.

Table 2: STEM jobs in the sample with explicit discipline requirements

STEM job =	*% STEM disciplines >50*		*% STEM disciplines = 100*	
	% of jobs that are STEM	% of STEM jobs belonging to	% of jobs that are STEM	% of STEM jobs belonging to
STEM occupations	81.64	69.46	78.45	70.63
Non-STEM occupations	24.11	30.54	21.92	29.37

Notes: Based on the sample of 3957387 vacancies with explicit discipline requirements and an occupation identifier. 1869128 STEM jobs, 1590254 jobs in STEM occupations.

Classifying a job as STEM if the percentage of STEM discipline requirements is above 50, Table 2 shows that over 30% of STEM jobs belong to Non-STEM occupations.[16] Restricting the definition to 100% STEM discipline requirements slightly lowers this percentage (29.37%).

Although these results are based on only 12% of all UK vacancies, they constitute an important robustness check and a preview of some of our findings because the sample with explicit discipline requirements has a 0.81 occupational correlation with the complete set of postings at the most refined 4-digit SOC level.

In what follows, our goal will be to classify all UK vacancies as STEM or Non-STEM based on the keywords collected from their online postings.

[16]Given the lack of a consistent "official" classification of four-digit occupations into STEM and non-STEM, we decided to merge together the lists from several widely cited UK studies: UKCES [23], Mason [41], BIS [8] and Greenwood et al. [30], resulting in a list of 73 four-digit STEM occupations: 1121, 1123, 1136, 1137, 1255, 2111, 2112, 2113, 2119, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2139, 2141, 2142, 2150, 2212, 2216, 2321, 2431, 2432, 2461, 2462, 2463, 3111, 3112, 3113, 3114, 3115, 3116, 3119, 3121, 3122, 3123, 3131, 3132, 3217, 3218, 3567, 5211, 5212, 5213, 5214, 5215, 5216, 5221, 5222, 5223, 5224, 5231, 5232, 5233, 5241, 5242, 5243, 5244, 5245, 5249, 5314, 8143.

3. Identifying STEM Keywords and Jobs

Irrespective of the occupations to which they belong, we want to identify STEM jobs as those whose vacancy descriptions contain "STEM keywords" - knowledge and skills that are typically acquired through a STEM education, or software/technological devices/job tasks that require and apply STEM knowledge & skills. Intuitively, recruiters employing STEM keywords when describing the job they want to fill and the candidate that they are looking for will be much more likely to seek to hire people with a STEM education even if they do not explicitly post a STEM discipline requirement.

Our approach consists of two steps: in Subsection 3.1, we identify STEM keywords using a method that we call "context mapping". The key idea in "context mapping", illustrated in Figure 1.1, is to classify keywords based on their "steminess" - the percentage of STEM discipline requirements with which the keywords appear in the sample where we observe both keywords and discipline requirements. Subsection 3.2 then proposes and evaluates several different ways of employing the steminess of the keywords found in an online vacancy posting to classify it as STEM or non-STEM, as well as estimate the probability that its recruiter looks for a STEM graduate.

3.1. **STEM keywords.**

The classification problem here is very simple: the BGT taxonomy contains 11,182 distinct keywords and we want to label as "STEM" those which correspond to knowledge and skills that are typically acquired through a STEM education, or software/technological devices/job tasks for which a STEM background is typically required in the labour market.

In theory, we could inspect all the keywords one by one and manually select those that seem to be STEM. In practice, this exercise is infeasible because of the thousands of technical terms, which may or may not be related to STEM, and that would require expert knowledge in order to be correctly classified, e.g.:

"Leachate Management", "Olas", "Step 7 PLC", "NASH", "Antifungal", "800-53"...

Even reading about these terms still leaves a lot of uncertainty and subjectivity in deciding on how to classify them. By contrast, this classification decision would be obvious to the recruiters employing these terms in their job descriptions since they should not only have a precise understanding of what these technical terms mean but also know the education background that successful job applicants for their advertised positions typically possess.

Luckily, 85.55% of all the BGT taxonomy (9566 keywords) ever appears in the subsample of vacancies with explicit discipline requirements (cf. Panel C, Table 1). Moreover, as shown in Fig. 2.1, for a vacancy selected at random from this sample, there is a roughly equal chance of finding a STEM or a non-STEM discipline requirement. Hence, a simple strategy, illustrated

in Figure 1.1, is to separate the 9566 "classifiable" keywords into STEM, Neutral and Non-STEM depending on the discipline "contexts" in which they appear. Intuitively, a proper STEM skill, knowledge, task should rarely appear together with a non-STEM degree because it requires a proper STEM education and a STEM qualification. Similarly, non-STEM skills (e.g. *"Cooking"*), knowledge (e.g. *"French"*), tasks (e.g. *"Account Reconciliation"*) would rarely appear in STEM contexts since they require a non-STEM education. At the same time, *"Communication skills", "Leadership", "Research", "Presentation skills"* are neither STEM, nor non-STEM specific skills, and therefore should not appear more often in vacancy descriptions of jobs requiring a STEM education than those requiring a non-STEM one. These are the "neutral" keywords.

Figure 1.1 shows some concrete examples: 95% of all disciplines with which the keyword *"C++"* appears are STEM. By contrast, *"English"* appears with STEM discipline requirements less than 30% of the time.

Let us refer to the percentage of STEM discipline requirements with which a keyword appears in the sample with both keywords and discipline requirements as its "steminess".[17]

After computing the steminess of all keywords, clustering techniques can be used to separate them into STEM, Neutral, and non-STEM, then further disentangle the STEM domain to which a STEM keyword is most likely to be related.

An important implicit assumption behind our strategy is that the subsample used to classify the keywords has the same underlying properties as the sample of all UK vacancies. As shown at the bottom of Panel C in Table 1, this seems to be the case since there is a 0.94 correlation between the frequency of posting a given keyword in the subsample with explicit discipline requirements and the sample of all postings. The correlations between the occupational and geographic distributions in the two samples are also very high: 0.81 and 0.99.

The insert on the next page shows the detailed steps of our strategy. We call it "context mapping" because the idea comes from Ethnography - the study of people and cultures. Ethnographers often seek to understand human behaviour by investigating "the environment in which the behaviour under study takes place", i.e. creating a "context mapping".[18] In our case, to understand whether a keyword should be classified as STEM, neutral or non-STEM, we look at whether the keyword appears more often with explicit STEM education requirements than with non-STEM ones, i.e. record the distribution of STEM vs. non-STEM discipline "contexts" in which the keyword appears.

In **Step 1**, we simply record, for any vacancy j that belongs to the sample with both keywords and explicit discipline requirements ($\mathcal{V}^D \cap \mathcal{V}^K$), the distribution of j's disciplines

[17]Credit to Rob Valletta for coining this term at the IZA Workshop. Also, we do not use "stemness" because it already has a precise definition in cytology (the study of cells).

[18]cf. http://www.ethnographic-research.com/ethnography/some-particular-methods/context-mapping/

Algorithm 1 *Context Mapping and Clustering*

Notation: Let $\mathcal{V} = \{j\}$ denote the set of vacancies (empty jobs), $\mathscr{K} = \{k\}$ the set of keywords, $\mathscr{D} = \{d\}$ the set of disciplines.

Vacancy j's online description contains keywords $\mathscr{K}_j$ and discipline requirements $\mathscr{D}_j$.

Define $\mathcal{V}^D \subset \mathcal{V}$ as the subset of vacancies that post at least one discipline requirement:

$$\mathcal{V}^D \coloneqq \{j | \mathscr{D}_j \neq \emptyset\}$$

Similarly, $\mathcal{V}^K \subset \mathcal{V}$ the vacancies with at least one keyword:

$$\mathcal{V}^K \coloneqq \{j | \mathscr{K}_j \neq \emptyset\}$$

Let $\mathscr{C} = \{C_1, ..., C_7\}$, with C_1 = Biology, C_2 = Physics, C_3 = Computer Sciences, C_4 = Technology, C_5 = Engineering, C_6 = Mathematics, and C_7 = Non-STEM.

Step 1: For all $j \in \mathcal{V}^D \cap \mathcal{V}^K$, record the distribution of j's discipline requirements over $\mathscr{C}$ as $\mathbf{c}_j = (c_{j,1}, ..., c_{j,7})$ with:

$$c_{j,p} = \frac{1}{|\mathscr{D}_j|} \sum_{i \in \mathscr{D}_j} \mathcal{I}(d_{j,i} \in C_p)$$

where $p = 1, ..., 7$, $\mathcal{I}(.)$ is the indicator function and $|.|$ denotes the cardinality of a set.

Step 2: Consider the set of keywords $\mathscr{K}^C \subset \mathscr{K}$ such that:

$$\mathscr{K}^C \coloneqq \{k \in \mathscr{K}_j | j \in \mathcal{V}^D \cap \mathcal{V}^K\}$$

For any $k \in \mathscr{K}^C$, let $\mathcal{V}_k \subset \mathcal{V}^D \cap \mathcal{V}^K$ be the subset of vacancies with discipline requirements that post k:

$$\mathcal{V}_k \coloneqq \{j \in \mathcal{V}^D \cap \mathcal{V}^K | k \in \mathscr{K}_j\}$$

Call $\mathcal{V}_k$ the "contexts" in which k appears and create a context mapping for k by taking the average distribution of disciplines in $\mathcal{V}_k$:

$$\mathbf{x}_k = (x_{k,1}, ..., x_{k,7}) \text{ with } x_{k,p} = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} c_{j,p}$$

The steminess of keyword k is defined as $steminess_k = 1 - x_{k,7}$.

Step 3: Classify the "classifiable" keywords $\mathscr{K}^C$ into three clusters $\mathscr{G}_1 = \{G_1, G_2, G_3\}$ with G_1 = STEM, G_2 = Neutral and G_3 = Non-STEM by minimizing:

$$\arg\min_{G_l} \sum_{l=1}^{3} \sum_{k \in G_l} (steminess_k - \overline{steminess}_l)^2$$

where $\overline{steminess}_l = \frac{1}{|G_l|} \sum_{k \in G_l} steminess_k$. The optimal parition is found using the algorithm described in Hartigan and Wong [31] with initial centroids selected as 0 (Non-STEM), 0.5 (Neutral) and 1 (STEM).

Step 4: Let $\mathscr{K}^{STEM}$ be the keywords identified in Step 3 as belonging to the STEM cluster. Classify $\mathscr{K}^{STEM}$ into six clusters $\mathscr{G}_2 = \{G_1, ..., G_6\}$ where $\mathscr{G}_2$ are the six STEM domains, e.g. G_1= Biology,..., G_6= Mathematics, by minimizing:

$$\arg\min_{G_l} \sum_{l=1}^{6} \sum_{k \in G_l} \sum_{p=1}^{7} (x_{k,p} - \bar{x}_{l,p})^2$$

The solution is found as in Step 3 but with initial centroids selected as $[\mathbf{I}_6; \mathbf{0}]$ with $\mathbf{I}_6$ being the 6×6 identity matrix.

over the six STEM domains and the non-STEM one in the vector $\mathbf{c}_j$. This step is necessary because 30% of vacancies post multiple disciplines. We then focus on the 9566 keywords $\mathscr{K}^C$ that ever appear in $\mathcal{V}^D \cap \mathcal{V}^K$ - the "classifiable" keywords. Whenever a keyword appears in a vacancy with discipline requirements, it appears in a "context" in which the distribution of disciplines over the STEM domains and the non-STEM one is given by $\mathbf{c}_j$. **Step 2** records the average distribution of disciplines among all the contexts in which k appears as $\mathbf{x}_k$. The steminess of a keyword is simply the proportion of STEM domains in $\mathbf{x}_k$. **Steps 3** and **4** implement a $K-$means clustering where we specify both the number of centers and their initial locations. In **Step 3**, we use the steminess of the keywords to partition them into STEM, neutral and non-STEM. The initial centroids are therefore 0, 0.5, and 1 corresponding to 0% STEM (Non-STEM cluster), 50% STEM (Neutral cluster) and 100% STEM (STEM cluster). **Step 4** classifies the STEM keywords into different STEM domains. The six initial centroids allocate 100% to each of the STEM domains.[19]

Figure 3.1 shows examples of randomly selected keywords from the resulting clusters. The method does not claim to be perfect. Nevertheless, "context mapping" does have the advantage of systematically classifying over 85% of all the BGT taxonomy, including many technical terms. More importantly, as Fig. 3.1 and further manual checks suggest, the resulting classification does seem fairly plausible.

For instance, *"Step 7 PLC"* is classified into the Technology cluster because it is an "engineering system in industrial automation".[20] *"NASH"* has nothing to do either with John Nash, or with STEM, or with non-STEM; it is the acronym for either "Non Alcoholic Steato Hepatitis", or "News About Software Hardware", or "Nashville"... Given this ambiguity, *"NASH"* cannot help us understand whether or not a job requires STEM knowledge and skills, hence the algorithm correctly classifies it as a neutral keyword. *"800-53"* is allocated to the Computer Sciences cluster since the "NIST Special Publication 800-53" is a catalog of security controls for federal information systems in the US. It is highly probable that people who would be referring to this publication in their jobs would also be required to understand how information systems work and are secured - knowledge that can be acquired through a degree like "Computer and Information Systems Security/Information Assurance" (cf. Table 16 in the Appendix).

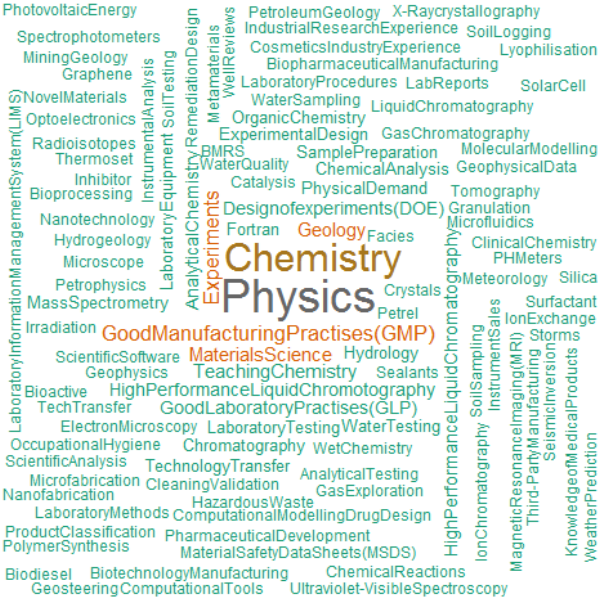
[19]Usually, in K-means clustering, the number of clusters is unknown. Researchers "try several different choices, and look for the one with the most useful or interpretable solution" (James et al. [34], chapter 10). Moreover, given a number of clusters, their initial locations (the centroids) are picked randomly and the resulting partition depends on this initial random selection. In our case both problems are avoided since the choices of the number of clusters and their locations are dictated by the type of information that we wish to extract. However, future research could explore more refined clustering or even other approaches: "with these methods, there is no single right answer—any solution that exposes some interesting aspects of the data should be considered."[34] Similarly, while the objective function in K-means clustering is the residual sum of squares, it would certainly be possible to try different criteria.

[20]http://w3.siemens.com/mcms/simatic-controller-software/en/pages/default.aspx

Figure 3.1 *Examples of STEM (by cluster), non-STEM and neutral keywords*

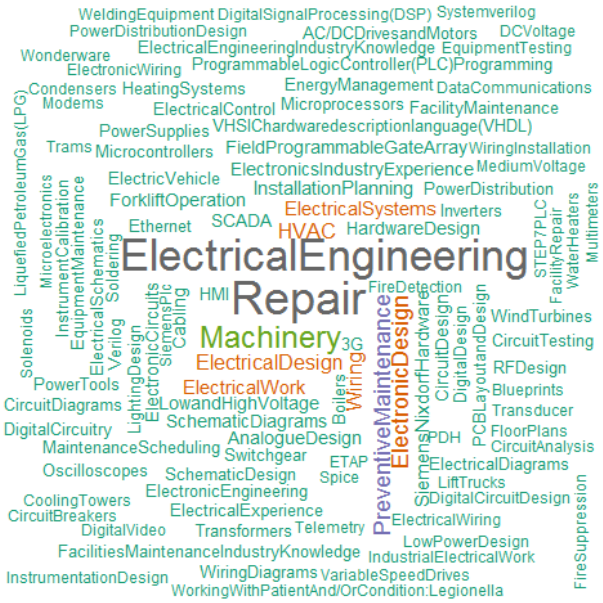
Biological & Biomedical Sciences

Physical Sciences



Computer Sciences

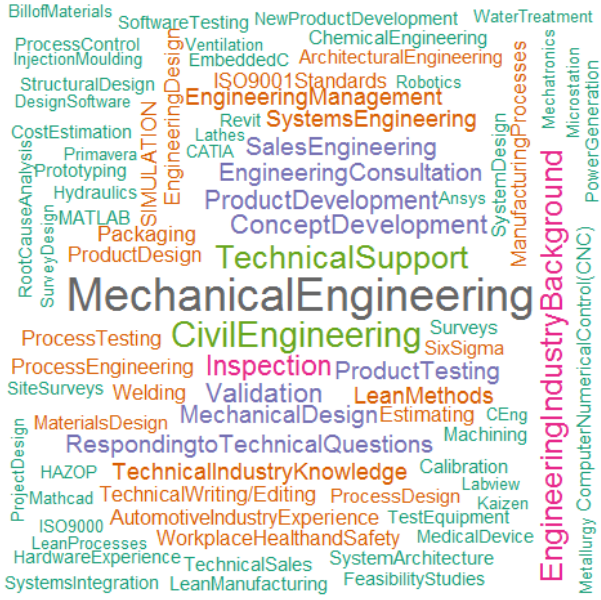
Technology



Notes: Continued on next page.

Note that keywords like *"Mathematics"*, *"Computer Skills"*, *"Problem Solving"* all appear in the neutral cluster. This is precisely because within the UK education system, such skills are not exclusively taught in STEM tracks. For instance, *"Mathematics"* on its own is often mentioned as a general basic skill requirement by many different recruiters looking for STEM and non-STEM graduates alike. It seems that a recruiter looking specifically for a

Engineering



Mathematics & Statistics

Non-STEM

Neutral

Notes: Random samples of around 100 distinct keywords collected from UK online vacancies and classified using context mapping and clustering. Size and color are by frequency of being posted. Figures created using *wordcloud* R package by Fellows [22].

Mathematics/Statistics graduate, would use much more precise keywords like *"Mathematical Modelling", "Statistics",* or technical terms, e.g. *"Chi-squared Automatic Interaction Detection (CHAID)*", *"Stochastic Optimisation*", etc. Indeed, the steminess of *"Mathematics"* is only 0.395, while it rises to 0.738 for *"Statistics"* and 0.892 for *"Mathematical Modelling".*

Table 3: STEM, Neutral and Non-STEM clusters

Cluster	*Steminess*			*No. Keywords*
	Mean	Median	Min	
STEM	0.89	0.91	0.69	3685
Neutral	0.49	0.50	0.29	2491
Non-STEM	0.10	0.08	0.00	3390

Notes: Summary statistics from the classification of 9566 keywords into STEM, Neutral and Non-STEM clusters.

Table 4: STEM domains clusters (STEM keywords only)

Cluster	*Average distribution of disciplines*							*No. Keywords*
	Biology	Computer	Engineering	Maths	Physics	Technology	Non-STEM	
Biology	0.73	0.02	0.05	0.03	0.05	0.01	0.11	754
Computer	0.01	0.53	0.23	0.05	0.02	0.03	0.13	639
Engineering	0.01	0.02	0.71	0.02	0.03	0.10	0.10	1266
Maths	0.07	0.12	0.12	0.49	0.05	0.02	0.13	152
Physics	0.12	0.02	0.22	0.03	0.45	0.05	0.11	372
Technology	0.01	0.02	0.30	0.01	0.02	0.55	0.09	502

Notes: Summary statistics from the classification of 3685 STEM keywords into six STEM domains.

BGT themselves classify 822 keywords as "Software and Programming". However, some of the software included in this category could be relatively easily learned/operated with no STEM background, e.g. *"Microsoft Excel"*, enterprise software like *"Oracle Human Resources"*, etc., or do not have much to do with a STEM education, e.g. *"Flickr"*, *"LinkedIn"*, etc. Context mapping classifies these keywords as either neutral or even non-STEM and clearly separates them from software and programming that do require advanced STEM knowledge and skills, e.g. *"Microsoft C#"*, *"UNIX Administration"*. Interestingly, among STEM software, statistical packages like *"SAS"* and *"R"* are assigned to the Mathematics & Statistics cluster because they mainly require knowledge of statistical analysis rather than very advanced computer programming skills. Other types of statistical software like *"Stata"* and *"E-Views"* are assigned to the neutral cluster because they are not more often taught in STEM disciplines than non-STEM ones. Hence, if such software were the only requirement the recruiter had, he would not be seeking STEM graduates with a higher probability than non-STEM ones.

Tables 3 and 4 provide further details on the distribution of discipline requirements within each cluster identified. For instance, the mean, median and min steminess of STEM keywords are 0.89, 0.91, and 0.69 respectively, while they are only 0.10, 0.08, and 0 for Non-STEM keywords respectively. Table 4 suggests that the Biology cluster is the best identified and most coherent with a 73% average loading on the Biological & Biomedical Sciences for the

Table 5: Classified vs. Unclassified Keywords

	2012	2013	2014	2015	2016	Total
% of Classified Keywords in a posting with >= 1 Keyword:						
Mean	99.99	99.99	99.99	99.98	99.98	99.99
Median	100	100	100	100	100	100
Number of Vacancies per Unclassified Keyword:						
Mean	4.20	4.34	3.75	9.26	6.05	13.57
Median	2	2	2	3	3	5
Number of Vacancies per Classified Keyword:						
Mean	3923.93	4373.20	3924.09	5071.65	3585.40	19264.55
Median	77	90	80	102	78	322

Notes: The classified keywords correspond to the 85.55% of the BGT taxonomy that ever appear in the sample with explicit disciplines and can therefore be classified using Algorithm 1.

754 keywords belonging to it. The Mathematics & Statistics cluster is the worst identified with only a 49% average loading on Mathematics & Statistics.

Note that although overall 85.55% of the BGT taxonomy are classified through Algorithm 1, the percentage of classified keywords in any given year actually ranges between 90.31% for 2015 and 94.03% for 2012 (cf. Table 1, Panel C). More importantly, Table 5 shows that, on average, 99.99% of all keywords posted in a vacancy with at least one are classified. A median vacancy has all 100% of its keywords classified. This happens because the unclassified keywords are precisely those that are posted least frequently: within the total sample of 33 million postings, the mean and median unclassified keywords appear respectively in 13.57 and 5 job ads, whereas for classified keywords the numbers are 19264.55 and 322 respectively.

3.2. **STEM jobs.**

3.2.1. *Steminess-based approaches.* Having classified individual keywords in the previous subsection, we now turn to the classification of jobs. And, since in our data jobs are nothing more than *sets* of keywords, e.g.:

"Training Programmes - Decision Making - Rugby"

classifying them is equivalent to labelling sets of keywords as STEM or non-STEM.

Perhaps the simplest way of doing this is to label those sets that contain at least one STEM keyword as STEM and the rest as non-STEM. Intuitively, since we identified STEM keywords as the skills and knowledge that are typically taught within STEM disciplines, or software/tools/technological devices/job tasks that apply STEM knowledge and skills, the presence of a STEM keyword in the vacancy description could well serve as an indicator for the fact that its recruiter is going to look preferably for someone with a STEM education.

Table 6: Vacancies classification, In-sample performance

Model		*% Correctly classified*	*% non-STEM misclas. into STEM*	*% STEM misclas. into non-STEM*	*Correlation with % STEM requirements*
			Panel A: Direct Methods		
(1)	STEM Keyword	84.04	23.27	8.44	0.665
(2)	Average Steminess	89.21	9.70	11.92	0.762
(3)	Weighted Av. Steminess	89.20	9.70	11.92	0.762
(4)	Naive Bayes	89.38	9.26	12.02	0.787
	Predictors		*Panel B: Logistic regressions*		
(5)	Mean Steminess	89.17	9.41	12.29	0.799
(6)	% STEM Keywords	87.02	7.87	18.25	0.754
(7)	Median Steminess	86.57	8.92	18.07	0.760
(8)	Max Steminess	85.83	17.31	10.93	0.740
(9)	Mean + % STEM	89.18	9.40	12.28	0.799
(10)	Mean + Median	89.32	9.70	11.70	0.802
(11)	Mean + Max	89.48	9.63	11.43	0.803
(12)	Mean + % STEM + Median + Max	89.47	10.02	11.06	0.804
			Panel C: Including Job Titles		
(13)	Naive Bayes	90.80	8.20	10.23	0.809
(14)	Mean + Max reg.	90.82	8.64	9.74	0.829

Notes: First three columns based on the sample of 3,554,318 vacancies with keywords and non-mixed discipline requirements. The correlation column employs the whole training sample (3,921,917 vacancies). In % STEM keywords we only consider classified ones. Weighted Average Steminess assigns a weight of 1 to any keyword that has been defined using at least 50 vacancies, then a weight of 0.5 + (No. vacancies/100) to all those that have been classified with less. All regression models (Panel B) include a constant and are estimated using a logit link function on the sample with non-mixed discipline requirements. The dependant variable is a dummy variable equal to 1 if all the discipline requirements are STEM and 0 if they are all non-STEM. Including job titles (Panel C) increases the training sample by 49,891 vacancies.

How well would this simple strategy work if implemented to recognize STEM and non-STEM jobs within the sample where discipline requirements are posted explicitly, i.e. the truth is known?

To address this question, we can create a so-called "confusion matrix":

	True outcome	
Prediction	Non-STEM disciplines	STEM disciplines
Non-STEM job	*Correct classification*	*Misclassified into Non-STEM*
STEM job	*Misclassified into STEM*	*Correct classification*

We classify jobs correctly if we predict STEM when the disciplines posted are indeed STEM and non-STEM when the explicit discipline requirements are also non-STEM. If our strategy predicts non-STEM (STEM) whereas the actual disciplines required are STEM (non-STEM), we have misclassified the job into non-STEM (STEM).

Hence, three indicators that tell us how well our strategy is at classifying jobs are: the percentage of jobs classified correctly, the % of non-STEM jobs misclassified into STEM and the % of STEM jobs misclassified as non-STEM.

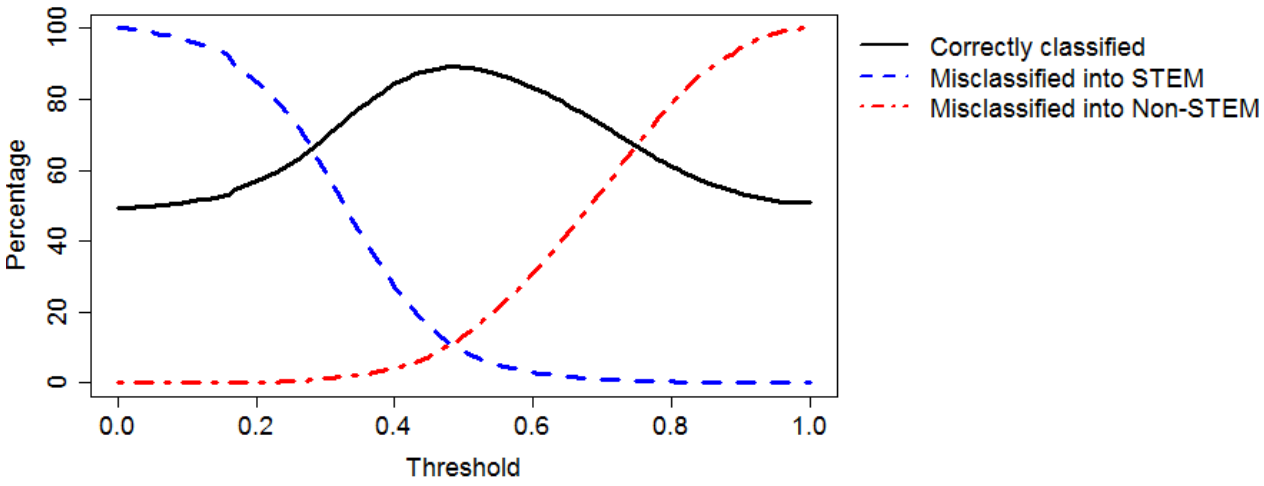
To avoid ambiguity, we focus on the subsample with non-mixed discipline requirements (i.e. either all STEM or all non-STEM) when computing the correct classification and misclassification rates in Table 6. To gauge the performance of our classifier on the sample with both mixed and non-mixed requirements, the last column of Table 6 shows the correlation of the predicted outcome with the % of STEM discipline requirements. Also note that the tests conducted in Table 6 are *in-sample* because the sample with explicit discipline requirements and keywords ($\mathcal{V}^D \cap \mathcal{V}^K$) used to evaluate our jobs classification strategies is the same sample that we used in the previous section to compute steminess and classify keywords. We implement *out-of-sample* tests in the following subsection.

The first proposed strategy corresponds to model (1). It classifies over 84% of vacancies correctly. Disagreggating the 16% error rate, the next two columns of the Table show that the "STEM Keyword" strategy misclassifies over 23% of non-STEM jobs into STEM, but has a much lower misclassification rate for STEM vacancies into non-STEM: only 8.44%. The relatively high misclassification rate into STEM occurs both because our classification of keywords is imperfect, but also because the meaning of a given keyword may be nuanced by the other keywords that appear with it in the job's description. In order to improve our correct classification rate, we therefore need an approach that somehow incorporates together all the keywords in the set that we want to label.

A direct approach here is to take the average steminess of all keywords in the job's description, then classify it as STEM if this average is above a certain threshold. Figure 3.3 shows that the correct classification rate peaks at 89.21% for a threshold of $\geq$0.49. Model (2) in Table 6 employs this optimal threshold. Misclassification into STEM drops substantially, from 23% to 9.7%. However, the misclassification into non-STEM rises by 3.5% pts. Note that average steminess (model 2) performs better than an indicator for the presence of at least one STEM keyword (model 1) not only in terms of overall correct classification, but also in terms of correlation with the % of STEM discipline requirements: 0.762 vs. 0.665.

An important concern is that the steminess of different keywords is computed using samples of different sizes with a median of 67 postings (cf. Panel C of Table 1). Taking a plain average gives equal weight to all keywords in the job description. On the one hand, down-weighting keywords that are defined using smaller sets could improve accuracy because their

Figure 3.3 *Using average steminess above a certain threshold to classify jobs as STEM*



Notes: The correct classification rate peaks at 89.21% for a threshold of an average steminess greater than or equal to 0.49.

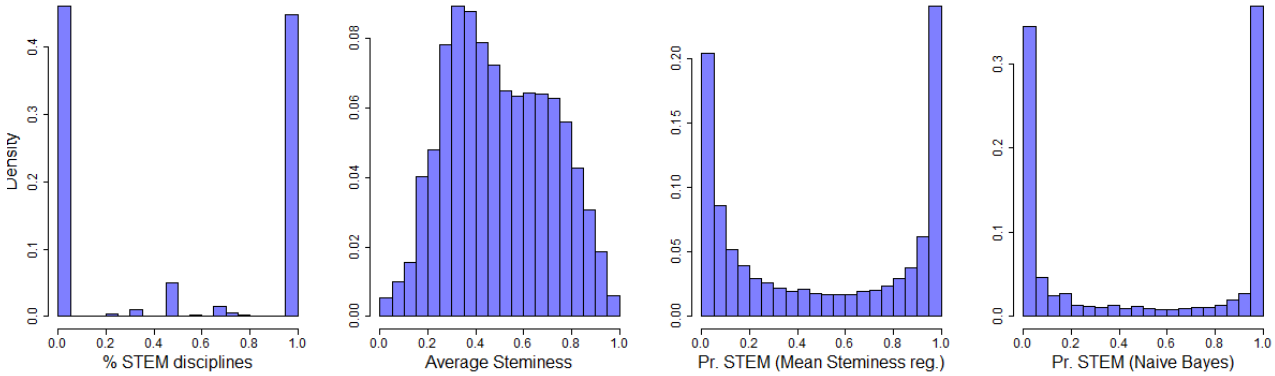
steminess is estimated less precisely. On the other hand, however, these keywords often correspond to some technical STEM terms and down-weighting them could make us believe that the average steminess of the job description is lower than it actually is. We tried several different weighting schemes. Overall, results are not very sensitive to the precise weighting. If anything, accuracy goes slightly down, suggesting that the technical terms argument may be more important than the precision one. For instance, in model (3) we assign a weight of 1 if a keyword's steminess is computed using at least 50 vacancies. Otherwise the weight is $0.5 + |\mathcal{V}_k|/100$. These weights are then normalized by the total weights' sum before taking the weighted average.

Although the simple unweighted average steminess performs surprisingly well with an almost 90% correct classification rate, there are several disadvantages of employing it. The first one can be seen from comparing the first two histograms in Fig. 3.4. The relatively high 0.762 correlation of average steminess with the true percentage of STEM degree requirements conceals the fact that the distributions in reality look quite different.

The second disadvantage is interpretation. The intuition is that a job description with a higher average steminess entails a more advanced requirement of STEM knowledge and skills. Its recruiter would therefore be more likely to want to hire a worker with a STEM education. Hence, ideally, we would like to use steminess to build an estimate of the probability of looking for a STEM graduate.

There are two ways of doing this. Firstly, instead of using mean steminess directly, we can employ it as the predictor in a regression that models the probability of requiring a STEM degree. In practice, to ensure that predicted probabilities lie between 0 and 1, we use a

Figure 3.4 *Comparison of classification strategies with the actual % of STEM disciplines required*



Notes: Based on the sample with keywords and discipline requirements.

logistic link function and estimate the following regression on the sample with non-mixed discipline requirements:[21]

$$\log\left(\frac{\Pr(\text{STEM} \mid steminess_j)}{\Pr(\text{Non-STEM} \mid steminess_j)}\right) = \alpha + \beta steminess_j \tag{3.1}$$

where $steminess_j = \frac{1}{|\mathscr{K}_j^C|}\sum_{k\in\mathscr{K}_j^C} steminess_k$ and $\mathscr{K}_j^C := \{\mathscr{K}_j \cap \mathscr{K}^C\}$ are the classified keywords in j's description. The dependent variable is an indicator for 100% STEM disciplines posted. We then use the estimated relationship to predict class probabilities for the complete sample (mixed and non-mixed disciplines) and classify jobs as STEM if $\Pr(\text{STEM} \mid steminess_j) > 0.5$.[22]

Note that the logistic regression allows for a non-linear relation between average steminess and the percentage of STEM degrees required which seems to fit the data better than a linear one since, even though the overall correct classification rate remains almost the same, the correlation between the predicted probabilities and the percentage of STEM disciplines posted is higher: 0.799 vs. 0.762. More importantly, the third histogram in Fig. 3.4 shows that the predicted probabilities match the distribution of the actual percentages of STEM disciplines required much better than raw average steminess.

Nothing prevents us from modelling the right hand side in eq.3.1 differently. For instance, instead of mean steminess, models (6), (7), and (8) use respectively the percentage of STEM keywords, the median and the maximum steminess as predictors. Interestingly, all these models achieve lower overall correct classification rates and correlations than mean steminess

[21]Using a probit link function instead of a logit one yields very similar results. All detailed regression results are available on request.

[22]This is equivalent to imposing a symmetric loss function on both the misclassification of non-STEM jobs into STEM and of STEM jobs into non-STEM (cf. Friedman et al. [35], chap. 2).

Table 7: Correlations between predictors

	% STEM Disciplines	Mean Steminess	% STEM Keywords	Median Steminess	Max Steminess
% STEM Disciplines	1.000	0.762	0.702	0.727	0.703
Mean Steminess	0.762	1.000	0.914	0.975	0.858
% STEM Keywords	0.702	0.914	1.000	0.903	0.741
Median Steminess	0.727	0.975	0.903	1.000	0.798
Max Steminess	0.703	0.858	0.741	0.798	1.000

Notes: Correlations based on whole training sample (3,921,917 vacancies). In % STEM Keywords, only classified ones considered.

on its own. Trying to combine them with the latter is also not very fruitful: overall precision does not rise by much in models (9), (10), (11) and (12). The reason is that, as shown in Table 7, all these predictors are highly correlated.

The regression that performs the best is the one with mean and max steminess as the predictors (model 11). Including maximum steminess is intuitively appealing because it helps ensure that we do not classify as STEM a vacancy description that just happens to only contain keywords with slightly above average steminess, but no keyword with really high steminess. Hence we keep model (11) as our preferred regression specification.

The second way of getting from steminess to the probability of requiring a STEM degree is to treat the steminess of each keyword k as the maximum likelihood estimate of $\Pr(\text{STEM} \mid k)$ - the probability of observing a STEM degree requirement conditional on observing k.

Let $\mathscr{K}_j^C = \{k_1, k_2, ..., k_{n_j}\}$, with n_j being the number of keywords collected from j's vacancy description. By Bayes' theorem:

$$\begin{aligned}\Pr(\text{STEM} \mid \mathscr{K}_j^C) &= \frac{\Pr(\text{STEM}, k_1, k_2, ..., k_{n_j})}{\Pr(k_1, k_2, ..., k_{n_j})} \\ &= \frac{\Pr(\text{STEM}) \cdot \Pr(k_1|\text{STEM}) \cdot \Pr(k_2|\text{STEM}, k_1) ... \Pr(k_{n_j}|\text{STEM}, k_1, k_2, ..., k_{n_j-1})}{\Pr(k_1, k_2, ..., k_{n_j})}\end{aligned}$$

Assuming that keywords are posted independently of each other, this expression simplifies to:

$$\Pr(\text{STEM} \mid \mathscr{K}_j^C) = \frac{\Pr(\text{STEM}) \cdot \prod_{k \in \mathscr{K}_j^C} \Pr(k|\text{STEM})}{\prod_{k \in \mathscr{K}_j^C} \Pr(k)} \tag{3.2}$$

$$= \frac{\prod_{k \in \mathscr{K}_j^C} \Pr(\text{STEM}|k)}{\Pr(\text{STEM})^{n_j - 1}} \tag{3.3}$$

where the last expression follows from observing that $\Pr(k|\text{STEM}) = \frac{\Pr(k) \cdot \Pr(\text{STEM}|k)}{\Pr(\text{STEM})}$.

Similarly, the probability of looking for a non-STEM graduate is

$$\Pr(\text{Non-STEM} \mid \mathscr{K}_j^C) = \frac{\prod_{k \in \mathscr{K}_j^C}(1 - \Pr(\text{STEM}|k))}{(1 - \Pr(\text{STEM}))^{n_j - 1}} \tag{3.4}$$

We can then classify a job as STEM if $\Pr(\text{STEM} \mid \mathscr{K}_j^C) > \Pr(\text{Non-STEM} \mid \mathscr{K}_j^C)$.

In text classification, this approach is known as the "multinomial Naive Bayes classifier", also sometimes called the "unigram language model" in the Information Retrieval literature (cf. Manning et al. [39], chapters 12 and 13). "Multinomial" because the ordering of the keywords does not matter, "naive" because of the naive assumption of independence.[23] Although this assumption is clearly wrong, in practice, there is simply no way of estimating more complex relationships between keywords given how sparsely they appear in the data and with each other.

Another practical issue is that because of multiplication, if any of the keywords in the description has a steminess (non-steminess) of zero, the predicted probability of looking for a STEM (non-STEM) graduate will be zero no matter the steminess (non-steminess) of the rest of the keywords. To remedy this issue, we simply need to smooth the steminess and non-steminess estimates so that they always lie in (0,1).

Remember from Algorithm 1 that steminess is computed as

$$steminess_k = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} c_{j,STEM} \tag{3.5}$$

where $c_{j,STEM}$ is the proportion of j's posted disciplines that are STEM and $\mathcal{V}_k$ is just the set of vacancies in which k appears in the sample with explicit discipline requirements.

Non-steminess is just $1 - steminess_k$:

[23]Strictly speaking, the standard implementation of Naive Bayes (NB) uses eq.3.2 and a similar expression for $\Pr(\text{Non-STEM} \mid \mathscr{K}_j^C)$ instead of equations 3.3 and 3.4. The reason we prefer the latter expressions is because they clearly show the link between steminess of keywords and the probability of looking for a STEM graduate, thereby empowering NB with our usual intuition that recruiters posting keywords with higher steminess look for STEM graduates with a higher probability. By contrast, the main input into the standard way of implementing NB is $\Pr(k|\text{STEM})$ interpreted as "a measure of how much evidence k contributes that STEM is the correct class" (Manning et al. [39]), i.e. the keywords are *not* of interest on their own, they are just a means of achieving the classification of jobs. The distinction is subtle but important since our logic is that the probability of looking for a STEM graduate and therefore of being classified as a STEM job is the direct consequence of the level of STEM skills and knowledge requirements implied by the keywords posted in the description (keyword steminess), i.e. the keywords are of primary importance.
In any case, we implemented both approaches to confirm that they give the same results which led us to realize that there is also a small "computational" advantage of implementing NB in the way we propose. Keywords appear in very few vacancies. Hence $\Pr(k|\text{STEM})$ are much smaller objects than $\Pr(\text{STEM}|k)$. For example, $\Pr(C++|\text{STEM}) = 0.00598$, while $\Pr(\text{STEM}|C++) = 0.95$. This is why the standard way of implementing NB often leads to a floating point underflow problem and is implemented by using a log transform. The log function is monotonic, hence the transform is not a problem if the only goal is classification. In our case, however, it is a problem because we also want the probability estimates. The floating point underflow problem is much less severe when NB is implemented using $\Pr(\text{STEM}|k)$.

$$non - steminess_k = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} c_{j,Non-STEM} \tag{3.6}$$

because $c_{j,STEM} + c_{j,Non-STEM} = 1$.

A simple way of smoothing is just to add a number to both steminess and non-steminess (Manning et al. [39], chap. 11). In our case, we can always let the keyword appear in at least one vacancy with perfectly mixed discipline requirements:

$$steminess_k = \frac{1}{|\mathcal{V}_k + 1|} \left\{ \sum_{j \in \mathcal{V}_k} c_{j,STEM} + 0.5 \right\} \tag{3.7}$$

$$non - steminess_k = \frac{1}{|\mathcal{V}_k + 1|} \left\{ \sum_{j \in \mathcal{V}_k} c_{j,Non-STEM} + 0.5 \right\} \tag{3.8}$$

Smoothing in this ways is like putting a uniform prior on whether the keyword appears with STEM or non-STEM disciplines and then letting the data update it. In any case, the correlation between smoothed and unsmoothed estimates for the 9566 classifiable keywords is over 0.98.

Note that because of smoothing and violations of the independence assumption, the probability estimates from equations 3.3 and 3.4 may be above one and not sum to one. However, we can simply normalize them as follows:

$$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) = \frac{\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)}{\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) + \widehat{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C)} \tag{3.9}$$

where:

$$\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) = \frac{\prod_{k \in \mathscr{K}_j^C} steminess_k}{\Pr(\text{STEM})^{n_j - 1}} \tag{3.10}$$

and

$$\widehat{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C) = \frac{\prod_{k \in \mathscr{K}_j^C} (non - steminess_k)}{(1 - \Pr(\text{STEM}))^{n_j - 1}} \tag{3.11}$$

and similarly for $\widetilde{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C)$. The correlations reported in Table 6 are with these normalized probability estimates.

As we can see the Naive Bayes approach (model 4) does quite well on our data: a correct classification rate of 89.38% and a correlation of 0.787. The last histogram in Fig. 3.4 suggests

that the pattern of predicted probabilities matches the distribution of the percentages of STEM discipline requirements quite well.

Another remarkable finding is that the correlation between STEM jobs identified using Naive Bayes and those identified using our preferred logistic regression with mean and max steminess as the predictors is 0.963. The correlation between their predicted probabilities is even higher: 0.968. This indicates that the two methods identify almost the same jobs as STEM and gives us confidence that a classification established with either of them will be accurate.[24]

3.2.2. *Out-of-sample performance & benchmarking against other ML algorithms.* At this point, the reader may have the following concerns about our strategy of classifying jobs into STEM and non-STEM:

(1) Endogeneity: the tests conducted in Table 6 are *in-sample* because the sample with explicit discipline requirements and keywords ($\mathcal{V}^D \cap \mathcal{V}^K$) used to evaluate our jobs classification strategies is the same sample that we use to compute steminess. How well do our preferred algorithms perform *out-of-sample,* i.e. on data that has not been used to estimate steminess? This is an important question since our ultimate goal is to classify all 33 million UK vacancies in our data, most of which do not have explicit discipline requirements, i.e. won't be used to estimate steminess for the final classification.
(2) Unclassified keywords: 15% of all keywords in the BGT taxonomy never appear with explicit discipline requirements and are therefore unclassified. How does this affect the performance of our algorithms?
(3) Steminess vs. keywords: our classification methods employ the steminess of all keywords in a vacancy description to either compute the mean and max steminess and use them as predictors in a logistic regression model, or to construct the probability estimate using Bayes formula and the naive assumption. A valid question is why not simply use the keywords directly instead of steminess to estimate the probability of looking for a STEM graduate?

We address these concerns by implementing and replicating 50 times the following experiment: we select 250,000 unique vacancies at random from the sample with non-mixed discipline requirements and keywords and split them into training (200,000 vacancies) and test (50,000) samples. To achieve a fair comparison, all methods discussed in this subsection are implemented on the same set of 50 randomly selected samples of 250,000 vacancies each.

[24]We tried an ensemble classifier which labelled a job as STEM only if both methods agreed on its classification. However, the performance of this ensemble classifier was not better in terms if overall classification: 89.41%. Hence, there seems to be no point in pursuing in this direction.

Table 8: Out-of-sample performance and Benchmarking

Model		*% Correctly classified*	*% Misclas. into STEM*	*% Misclas. into non-STEM*	*Computing Time (hh:mm:ss)*	*Computer Memory (Gigabytes)*	*% Failed*
(1)	Mean + Max reg.	89.53 [0.134]	9.71 [0.198]	11.26 [0.191]	00:05:35 [00:00:43]	4.70 [0.001]	0
(2)	Naive Bayes	89.60 [0.138]	9.22 [0.221]	11.62 [0.201]	00:05:44 [00:00:48]	4.54 [0.001]	0
(3)	Logistic Regression with Keywords	87.16 [0.176]	6.39 [0.332]	19.50 [0.562]	04:57:26 [00:44:20]	14.91 [0.046]	0
(4)	Linear Discriminant Analysis	89.945 [0.140]	7.770 [0.212]	12.407 [0.277]	08:31:57 [00:59:47]	95.79 [6.645]	36
(5)	Support Vector Machines	90.24 [0.128]	6.59 [0.211]	13.04 [0.237]	09:25:42 [00:51:54]	14.81 [0.705]	2
(6)	Tree	72.918 [0.410]	2.652 [6.578]	52.260 [6.725]	04:05:38 [00:36:51]	52.46 [0.490]	8
(7)	Boosting Tree	77.044 [1.763]	3.034 [1.047]	43.496 [4.425]	05:43:40 [01:00:04]	56.10 [3.308]	16
(8)	Bagging Tree						100
(9)	Random Forests						100
(10)	Neural Networks						100
(11)	k-Nearest Neighbours						100

Notes: Bootstrapped standard errors in brackets. Averages over 50 runs of the experiment shown. The same set of 50 randomly selected samples of 250,000 vacancies each (split into 200,000 vacancies for the training sample and 50,000 for the test one) was used to evaluate all methods. All R scripts were submitted to the same High Performance Computing cluster and the statistics presented here are those that were output by the system once the jobs had been completed. The *RTextTools* package (Boydstun et al. [10]) was used for the implementation of the standard classification methods. As discussed in [10], this package employs a set of optimized algorithms, in particular the *SparseM* package by Koenker and Ng [37]. The R code for the implementation of all the algorithms is available on request. Computing time corresponds to the user time which is the time spent on executing the script's code lines. "User time" is typically reported for algorithmic benchmarking and performance analytics because it does not count the "System time" - time spent by the system on opening the files (which in our case was 8 sec or less for the first two methods that employ steminess and between 33 sec and 3 min 45 sec for the standard algorithms).

The results are summarized in Table 8 which reports the average correct classification and misclassfication rates over all the replications and the bootstrapped standard errors in brackets. We now discuss in turn why this out-of-sample experiment addresses each point just identified:

Issue (1) is addressed directly since we are implementing *out-of-sample* tests. Each time, the 200,000 vacancies in the training sample are used to train the algorithm, e.g. for the first method, to estimate steminess for all keywords, then run the logistic regression with mean and max steminess as the predictors. The trained model is then used to predict the outcomes for the test sample of 50,000 vacancies. The statistics reported in Table 8 are based on the performance of our algorithms on these latter test vacancies only. It is reassuring to

see that both of our preferred methods perform as well out-of-sample as they did in-sample, with almost 90% correct classification rates.

For the second issue, note that out-of-sample tests recreate the situation of having a certain proportion of unclassified keywords and therefore allow us to gauge the extra degree of misclassification generated by not being able to classify all keywords. Concretely, in our experiment, the training samples contained an average of 6810 distinct keywords. The test samples had on average 5210 distinct keywords, of which an average 244 were undefined. On average, 49999 vacancies were classified each time (i.e. one of the vacancies could not be classified because none of its keywords could be defined). The extra misclassification introduced by not being able to define all keywords happens to be very small since the average percentages of vacancies classified correctly in-sample are only slightly higher than the out-of-sample ones shown in Table 8: 89.73% for the logistic regression with mean and max steminess as the predictors, and 89.78% for Naive Bayes.

To address issue (3), we implemented several standard machine learning algorithms, often employed for supervised text classification.[25] The one thing they have in common is that they use the keywords directly, i.e. their implementation starts with the creation of a so-called "document-term" matrix (more precisely a "vacancy-keyword" matrix in our case) whose elements are 0-1 vectors that record for each vacancy the keywords collected from its description. The idea is then to divide the keywords ("the input space") into a collection of regions labelled as STEM and non-STEM (cf. Friedman et al. [35], chapt. 4). The methods differ in how exactly this division is made. For instance, logistic regression with keywords as predictors (and regularized versions thereof) or linear discriminant analysis (LDA) have linear decision boundaries. In support vector machines (SVM), a non-linear hyperplane separates STEM and non-STEM regions, allowing for some misclassifications that we can control with a cost parameter. Tree-based methods are called so because they try to segment the input space into a number of non-overlapping regions through a set of splitting rules that can be summarized in a tree. Bagging, Boosting and Random Forests are just more complex variants of the plain tree, which involve producing multiple trees, then combining them in order to yield consensus predictions.

A remarkable finding in Table 8 is that a logistic regression with almost 7,000 distinct keywords as the predictors (model (3)) achieves a 2.4% pts. *lower* correct classification rate than a logistic regression with just *two* predictors: the mean and the max steminess (model (1)). Moreover, it is much more computationally intensive: when the keywords are used

[25]Gareth et al. [34] is an excellent introduction into statistical learning. Friedman et al. [35] provide a more advanced treatment of a similar set of topics. In terms of books specifically focused on text analysis, we refer the reader to Feldman and Sanger [21] and the fascinating book on information retrieval by Manning et al. [39].

directly in the logistic regression, the average run of the experiment takes almost 5 hours instead of a bit more than 5 minutes, and consumes over 10 more gigabytes.

Tree methods perform worse than our preferred algorithms. Although they misclassify very few non-STEM vacancies into STEM, this comes at a high price of mislabelling around half of STEM vacancies as non-STEM. The very large misclassification into non-STEM occurs because of the way in which trees work: they use the presence of a keyword in a vacancy description as a split condition. Hence, many STEM vacancies are mistakenly assigned to the non-STEM group simply because they contain certain keywords that also happen to be often found in non-STEM vacancies.

The only two methods that seem to perform slightly better than the steminess-based classification algorithms are LDA and SVM . However, this performance comes at a much higher computational cost: 8h32 and 95.8 gigabytes on average for LDA, 9h26 and 14.8 gigabytes for SVM. By contrast, our preferred methods take on average less than 6 minutes and less than 5 gigabytes in each replication of the experiment. We relied on the *RTextTools* package by Boydstun et al. [10] for the implementation of models (3) to (10). Although, this package employs a set of optimized algorithms, in particular those developed by Koenker and Ng [37] and contained in the *SparseM* package, there could certainly be more efficient ways of implementing the standard machine learning algorithms considered here both in R or other programming environments. Nonetheless, the computational complexity of these methods is well studied and documented (cf. Manning et al. [39] and Friedman et al. [35], as well as references therein). The problems become especially acute when the input space is high dimensional and sparse, which is precisely our case: as both the number of distinct keywords and vacancies grow, the "vacancy-keyword" matrix becomes increasingly sparse because even the median keyword appears in very few postings (less than 0.002% in the sample of vacancies with explicit discipline requirements and keywords, which is the sample on which the final classification method is trained). Note that regularization (e.g. Lasso, Ridge) here does not help for two reasons: the optimally selected penalty (though cross-validation) is close to zero. More importantly, even if we remove 50% of all keywords, we are still left with a very sparse matrix.

This "sparse sampling in high dimensions" is often referred to as the "curse of dimensionality" (Friedman et al. [35]) and is also the reason why many methods (models (8)-(11) in Table 8) simply fail. For instance, we tried kNN with different numbers of neighbours; however the method failed because in our data few vacancies have many overlapping keywords so that the nearest neighbours are numerous but not "close to the target point" (Friedman et al. [35]).

Indeed, a conceptually more important problem with using the keywords directly is that this approach treats all the thousands of distinct keywords as completely *separate* dimensions,

i.e. it does not allow a keyword like *"Budgeting"* to be closer to *"Budget Management"* than to *"Java"*.

Employing keyword steminess instead of using the keywords directly is like introducing one extra step in-between the keywords and the prediction about whether the job is STEM or not. However, this extra step solves all the problems. The vacancy-keyword matrix is not needed which saves a lot of computing power. The logistic regression problem is much simpler in model (1) than model (3) because the predictive relationship is built from just two continuous predictors (mean and max steminess) instead of several thousands of dummy variables. In terms of steminess, *"Budgeting"* (65.59%) is indeed much more similar to *"Budget Management"* (63.80%) than to *"Java"* (95.13%).

Finally, throughout this section, we spent a lot of effort building the intuition behind the concept of steminess, the context mapping method for classifying keywords and eventually the steminess-based classification methods for the jobs. By contrast, the intuition underlying most standard machine learning methods presented here seems less straightforward since many of them were developed with the only goal of yielding accurate predictions, not necessarily being used for inference (Gareth et al. [34]). They treat the keywords as simple features, with no interest in classifying them or understanding how and why they should or should not be associated with the probability of looking for a STEM graduate, while the precise mechanisms used to split these keywords so as to form predictions for the jobs remain a bit of "black boxes".

3.2.3. *Job Titles.* While 90% of vacancies have at least one keyword collected from their online description, the job title is available in 100% of the cases (cf. Table 1). Employing keywords from the job titles could therefore not only improve our classification accuracy, but, more importantly, should allow us to classify more vacancies.

Unlike the vacancy descriptions which are already in the form of sets of keywords in our data, the job titles appear as sentences, e.g.: *"Principal Civil Engineer", "Uk And Row Process Diagnostic Business Manager", "Nurse Advisor"...*

We therefore start by tokenizing them, i.e. "chopping character streams into tokens" (Manning et al. [39]). For instance, tokenizing:

"Uk And Row Process Diagnostic Business Manager"

gives the following set of keywords:

"Uk - And - Row - Process - Diagnostic - Business - Manager"

This produces a list of over 143,000 distinct keywords which contains a lot of noisy terms, e.g. "aaa". To reduce and clean it, we implement several natural language processing steps. Firstly, we match whatever we can with the keywords from the BGT taxonomy. Another advantage of doing this is to increase the number of vacancies in which a given keyword from the BGT taxonomy appears. For the remaining keywords, we only focus on those

Table 9: Including keywords from job titles

	2012	2013	2014	2015	2016
% with >= 1 Classified Keyword	99.92	99.79	99.83	99.85	99.82
No. of unique keywords	27025	28218	28485	29567	27599
Number of Keywords per Vacancy (conditional on at least one classified):					
Median	7	7	7	7	7
Mean	8.72	8.65	8.91	8.89	8.72

Notes: 2016 includes data up to (and excluding) August only. Classified Keywords include 9566 keywords from the BGT taxonomy and 20,265 tokens from the job titles.

appearing in at least 10 postings. These simple steps already remove a lot of idyosincratic noise and reduce the list down to 20,615 unique keywords. We then remove punctuation marks, numbers, special characters, transform the tokens to lower space, and delete English stop words (e.g. "and", "I", "very", "after", etc.).[26]

For instance, the above title becomes:

"uk - row - process - diagnostic - business - manager"

We add the resulting tokens to the BGT taxonomy as extra features, so that the final classification of vacancies is based on 29,831 unique keywords - 9566 from the original BGT taxonomy and 20,265 from the job titles.

As shown in Panel C of Table 6, in-sample performance of our preferred classification methods jumps above 90%. However the real advantage of including keywords from job titles can be seen by comparing Tables 1 (Panel B) and 9. Now almost 100% of all vacancies have at least one classified keyword. The mean and median numbers of classified keywords per vacancy increase from to 5 and 6 to 7 and almost 9 respectively.

In what follows we use the Naive Bayes method with keywords from both the BGT taxonomy and the job titles to classify vacancies. The results employing the mean and max steminess regression for the classification of jobs are almost identical, since, as already discussed, both methods have an above 0.96 correlation for both the STEM jobs identified and the probability estimates.

4. STEM Jobs In the UK

Having designed and tested algorithms that classify both keywords and jobs into STEM and non-STEM with a 90% correct classification rate for the jobs, we can now finally start exploring our main questions of interest: what percentage of STEM jobs are in non-STEM

[26]Sanchez [46], Baayen [4] and Feinerer et al. [20] are excellent references on text processing and analysis in R. Natural language processing R packages used in this project include *stringi* (Gagolewski and Tartanus [28]), *stringr* (Wickham [49]), *tm* (Feinerer et al. [19]), *NLP* (Hornik [33]), and *quanteda* (Benoit [7]).

Table 10: STEM jobs vs. STEM occupations

	2012	2013	2014	2015	2016	Total
No. STEM jobs	1949791	2235445	1815294	2655532	1865435	10521497
No. STEM jobs in Non-STEM occ.	633578	798933	643232	914609	645961	3636313
No. STEM jobs in STEM occ.	1316213	1436512	1172062	1740923	1219474	6885184
No. jobs in STEM occ.	1580088	1764163	1495158	2146155	1500800	8486364
% of STEM jobs in						
STEM occupations	67.51	64.26	64.57	65.56	65.37	65.44
Non-STEM occupations	32.49	35.74	35.43	34.44	34.63	34.56
STEM density of						
STEM occ.	83.30	81.43	78.39	81.12	81.25	81.13
Non-STEM occ.	14.59	15.23	13.66	15.27	15.61	14.89

Notes: Based on the sample of vacancies with a UK SOC identifier (99.5% of all vacancies posted). For the list of STEM occupations cf. Footnote 16. 2016 includes data up to August only.

occupations? Are STEM jobs associated with higher wages? What, if anything, distinguishes STEM jobs in STEM vs. non-STEM occupations?

When documenting occupational and geographic distributions, we consider the following two indicators of STEM importance.

Let A be an occupation or a county:

(1)
$$\text{STEM density of } A = 100 \times \frac{\#(\text{STEM jobs in } A)}{\#(\text{jobs in } A)}$$

(2)
$$\%\ \text{STEM jobs in } A = 100 \times \frac{\#(\text{STEM jobs in } A)}{\#(\text{STEM jobs})}$$

While the percentage of STEM jobs simply describes how STEM jobs are distributed across occupations/counties, the STEM density measures the relative importance of STEM within an occupation/county. The higher the STEM density, the bigger the proportion of recruiters within this occupation/county that require STEM skills and knowledge.

4.1. **Occupational distribution.** The first goal of this paper is to go beyond STEM *occupations* and quantify the demand for STEM at the level of *jobs*. Table 10 therefore presents our main results which indicate that it is wrong to equate STEM demand with STEM occupations.

Firstly, the overall number of STEM jobs is larger than the number of jobs in STEM occupations. For instance, in 2015, focusing exclusively on STEM occupations leads to understimating the true demand for people with a STEM education by half a million employment opportunities.

Secondly, around 35% of all STEM jobs are in non-STEM occupations. Hence, the fact that over half of STEM graduates work in non-STEM occupations may be less problematic than often thought if most STEM graduates working in non-STEM occupations are actually in STEM jobs.

As expected, a much larger proportion of jobs within STEM occupations are STEM than within non-STEM ones: 81% vs. 15%. However, these aggregate numbers conceal an important amount of heterogeneity illustrated in Figure 4.1 which shows the distribution of STEM densities at the four-digit UK SOC level.

Table 15 in the Appendix contains the precise numbers for 2015. The third column in this table is a dummy indicator for whether or not the occupation is typically classified as STEM. Given the absence of a consistent "official" classification of four-digit occupations into STEM and non-STEM, we decided to merge together the lists from several widely cited UK studies: UKCES [23], Mason [41], BIS [8] and Greenwood et al.[30] (for the resulting full list of STEM occupations, cf. Footnote 16).

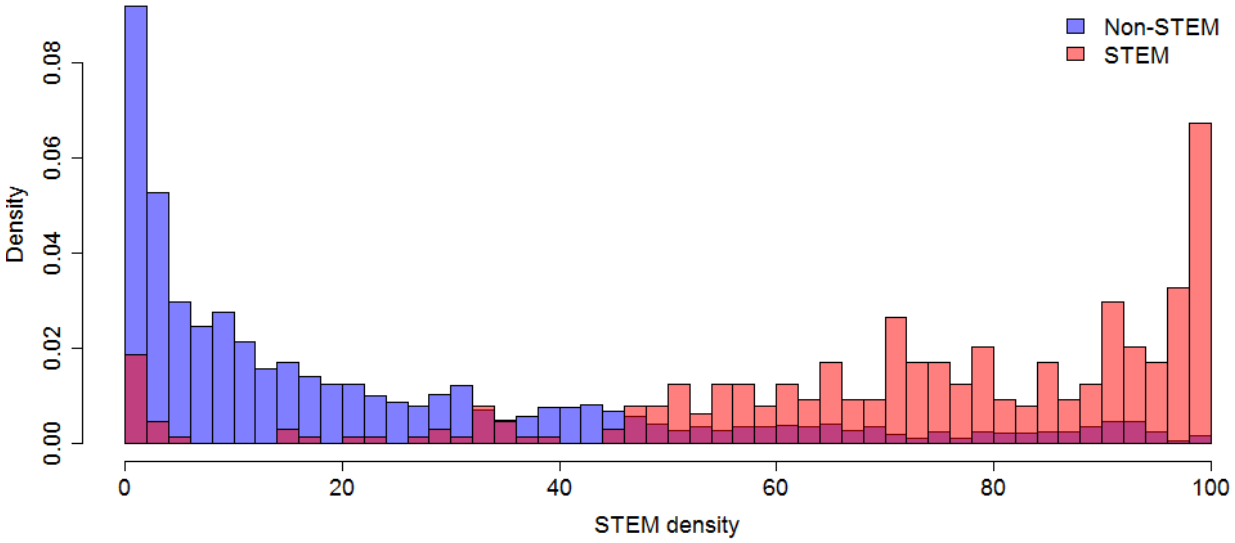
When interpreting the STEM densities of various occupations, it is important to remember that while we take the STEM acronym literally, some of these studies have a broader definition of STEM, which goes beyond Sciences, Technology, Engineering and Mathematics and also includes subjects like Medicine, Architecture, Environmental Studies, Psychology etc. Hence, in some cases, for instance *Pharmaceutical technicians (3217)*, we find a low STEM density in a STEM occupation precisely because of this broader STEM definition effect. In other cases, however, e.g. *Information technology and telecommunications directors (1136), Quality assurance and regulatory professionals (2462)*, the relatively low STEM density suggests that the occupation is less STEM intensive than typically thought.[27]

The list of non-STEM occupations with relatively high STEM densities is very diverse. For instance, in 2015, 46.84% of *Business, research and administrative professionals n.e.c.* jobs were identified as STEM. 45.62% of *Product, clothing and related designers,* and even 23.46% of *Artists* looked for STEM graduates. This finding recalls another passage from Matthew Sigelman's inspiring essay on "Why the STEM Gap is Bigger Than You Think" [47] where the opening quote also comes from: "the list [of job categories where employers demand coding skills] includes Artists and Designers, which once would have been considered the antithesis of STEM roles."

Perhaps surprisingly for the literature, where financial occupations are typically considered as the main non-STEM group poaching STEM graduates, none of them is actually top of the list in terms of STEM density. For instance, among the seven occupations defined as financial

[27]Another caveat to bear in mind is that there may be some misclassifications in our data because of imperfections in the collection process and/or errors in the online postings themselves. Moreover, as established in the previous section, our classification algorithm has a 90% correct classification rate, hence it misclassifies around 10% of jobs.

Figure 4.1 *STEM density of STEM and Non-STEM Occupations*



Notes: STEM density is the percentage of jobs within an occupation that are STEM. All years combined: an observation is a four-digit occupation-year STEM density.

in Chevalier [14], *Management consultants and business analysts* is the one with the highest percentage of STEM jobs in 2015: 25.33%, followed by *Financial and accounting technicians* with 11.67%. Only 7.59% of *Finance and investment analysts and advisers* specifically look for STEM graduates. The reason may be that, within the UK education system, the "numerical skills" for which financial occupations are thought to be seeking STEM graduates are actually also often transmitted to non-STEM graduates in, e.g., *Finance* or *Economics* degrees. Hence, although numerous jobs in financial occupations may end up being filled with STEM graduates, when posting their vacancy, not many financial recruiters actually describe the job as one that could only be undertaken by someone with a STEM education.

The main focus of this paper is on "high-level" STEM jobs - STEM jobs belonging to Managerial, Professional and Associate professional positions which typically require a university degree - because they constitute 74% of all STEM jobs (cf. fourth column of Table 11, occupation codes 11 - 35), but also because this is where the biggest expenses on STEM education are and where the STEM pipeline leakage is therefore most problematic.

However, Table 11, which compares the occupational distributions of STEM jobs vs. jobs in STEM occupations at the two-digit level of the UK SOC, suggests that many lower skill occupations with relatively high STEM densities are completely missed in the existing classifications of STEM occupations. Indeed, almost all four-digit occupations identified as STEM in the studies by BIS [8] and Mason [41] that investigate vocational STEM skills and apprenticeship training, belong to *Skilled Metal, Electrical and Electronic Trades* (cf. Table 11, fifth

Table 11: Occupational distribution of STEM jobs in 2015

Code	*Name*	*STEM density*	*% of STEM jobs in*	*% jobs in STEM occ.*
11	Corporate Managers and Directors	24.31	4.33	15.82
12	Other Managers and Proprietors	29.59	2.79	0.16
21	Science, Research, Engineering and Technology Professionals	85.26	39.73	99.77
22	Health Professionals	1.63	0.24	3.23
23	Teaching and Educational Professionals	2.99	0.3	0
24	Business, Media and Public Service Professionals	25.45	6.82	14.8
31	Science, Engineering and Technology Associate Professionals	76.12	11.46	100
32	Health and Social Care Associate Professionals	5.92	0.23	15.22
33	Protective Service Occupations	24.47	0.13	0
34	Culture, Media and Sports Occupations	15.73	0.98	0
35	Business and Public Service Associate Professionals	16.04	6.97	1.93
41	Administrative Occupations	5.93	1.27	0
42	Secretarial and related Occupations	4.05	0.28	0
51	Skilled Agricultural and related Trades	20.58	0.09	0
52	Skilled Metal, Electrical and Electronic Trades	89.79	9.23	94
53	Skilled Construction and Building Trades	61.92	2.29	24.33
54	Textiles, Printing and other Skilled Trades	8.45	0.57	0
61	Caring Personal Service Occupations	1.61	0.18	0
62	Leisure, Travel and related Personal Service Occupations	7.44	0.26	0
71	Sales Occupations	12.53	1.88	0
72	Customer Service Occupations	6.93	0.43	0
81	Process, Plant and Machine Operatives	60.38	4.74	0.21
82	Transport and Mobile Machine Drivers and Operatives	35.88	2.25	0
91	Elementary Trades and related Occupations	57.81	1.4	0
92	Elementary Administration and Service Occupations	12.21	1.16	0

Notes: Based on the sample of vacancies with a UK SOC identifier (99.5% of all vacancies posted). 2-digit UK SOC Classification.

column). However, our analysis suggests that STEM skills required in *Skilled Construction and Building Trades* and *Process, Plant and Machine Operatives* occupations should also receive more attention in future work since they represent together around 7% of STEM jobs and have STEM densities above 60%.

These findings echo a recent US study by Rothwell [45] who argues that: "previous reports on the STEM economy indicate that only highly educated professionals are capable of mastering and employing sophisticated knowledge in STEM fields. Classifying STEM jobs based on knowledge requirements, however, shows that 30 percent of today's high-STEM jobs are actually blue-collar positions. As defined here, blue-collar occupations include installation, maintenance, and repair, construction, production, protective services, transportation, farming, forestry, and fishing, building and grounds cleaning and maintenance, healthcare support, personal care, and food preparation."

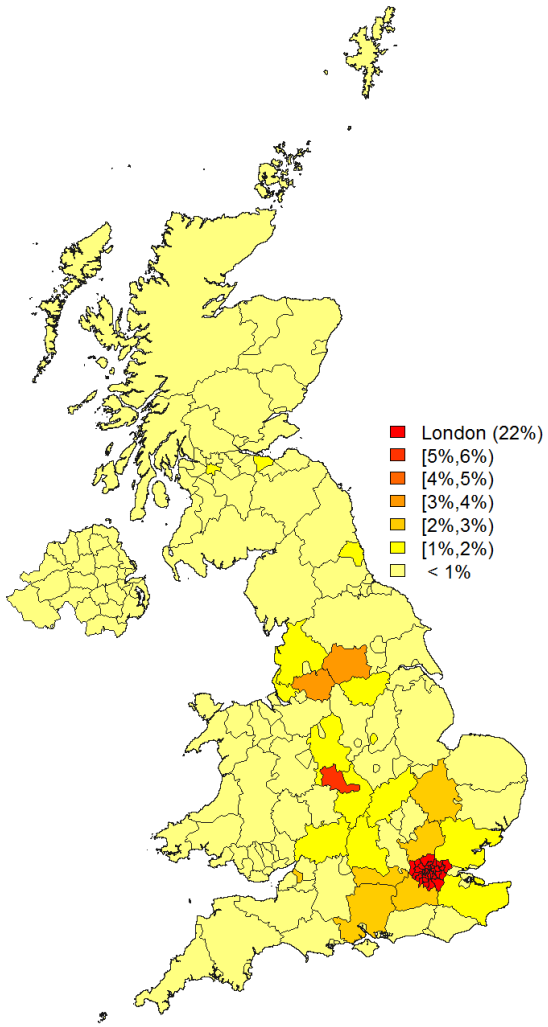
The reason why Rothwell identifies this category of STEM employment is because he uses a very different way of identifying STEM occupations, based on data from the O*NET (Occupational Information Network Data Collection Program) - a comprehensive database developed by the US Department of Labor, "which uses detailed surveys of workers in every occupation to thoroughly document their job characteristics and knowledge requirements." Rothwell focuses on O*NET Knowledge scales for Biology, Chemistry, Physics, Computers and Electronics, Engineering and Technology, and Mathematics. These scales are constructed by asking around 24 workers from each occupation to rate the level of knowledge required to do their job. For instance, the survey asks the worker: "What level of knowledge of Engineering and Technology is needed to perform your current job?" It then presents a 1-7 scale and provides examples of the kinds of knowledge that would score a 2, 4, and 6. Installing a door lock would rate a 2; designing a more stable grocery cart would rate a 4; and planning for the impact of weather in designing a bridge would rate a 6 (O*NET [24]). In some sense, our keywords-based approach of identifying STEM jobs is akin to surveying not workers as in O*NET, but employers, and this explains why our results also reflect all the "diversity and depth of the STEM economy".

4.2. **Spatial distribution.** Existing studies indicate that London is a "magnet of STEM workers at the expense of other parts of the country". For instance, Bosworth et al. [9] analyse commuting data and find that London has a net gain of 87,000 Core STEM workers, while the South East, East of England and East Midlands record substantial net losses.

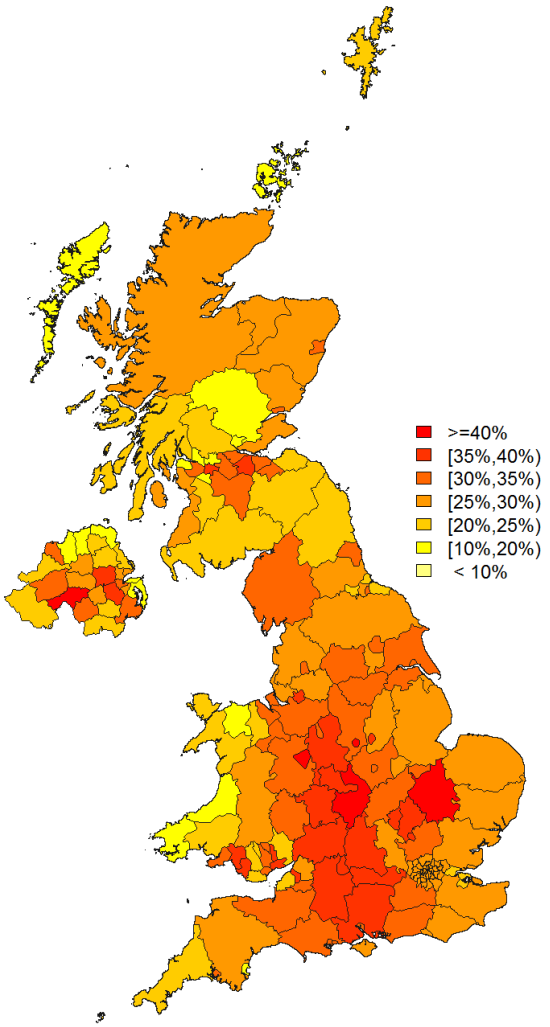
Although London is over-represented in the BGT sample relative to official employment data (cf. Table 18 in the Appendix), the first map in Figure 4.2 shows that it still has by far the greatest concentration of all STEM vacancies, explaining why it may be so attractive to STEM educated job seekers. In 2015, London concentrated 22% of all STEM vacancies with the next biggest demand for STEM knowledge and skills coming from West Midlands (which includes Birmingham and Coventry) with only 5.5% of STEM vacancies, followed by Greater Manchester (3.6%) and West Yorkshire (3.5%). Less than 3% of STEM vacancies were located in any other county.

Figure 4.2 *The geographical locations of STEM vacancies in 2015*

% of STEM jobs in each county



STEM density of each county



Notes: Based on the sample of 77.8% of all vacancies with County identifiers in 2015. London includes the 32 London boroughs and the City of London. STEM density is the % of jobs within a county that are classified as STEM.

In terms of STEM density (second map), the picture is less clear-cut. In 2015, London had a STEM density of 29.97%, while Cambridgeshire came top with 45.51%. Note that none of the counties had a STEM density below 10%, suggesting that at least some STEM knowledge and skills are required in every UK county.

Interviews with STEM employers, analysed in Bosworth et al. [9], reveal that some of them experience hiring difficulties "because their location is outside of London". Hence, the overall

message from previous studies and the spatial distribution of STEM vacancies analysed here could be that many STEM workers may move to London thinking that it would be easier for them to find a STEM job there since London concentrates over 20% of all STEM vacancies. This, however, induces shortages in some areas since most UK counties need at least a certain proportion of their workforce to possess STEM knowledge and skills.

4.3. **The wage premium for STEM.** To examine whether or not STEM jobs are associated with higher wages in the labour market, we run simple linear regressions like:

$$\log w_j = \alpha + \beta STEM_j + \gamma \boldsymbol{X}_j + \varepsilon_j \tag{4.1}$$

$$\log w_j = \alpha + \beta \widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) + \gamma \boldsymbol{X}_j + \varepsilon_j \tag{4.2}$$

where w_j is the hourly wage, $STEM_j$ is an indicator for whether the job is classified as STEM, $\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$ is the probability that the recruiter for vacancy j seeks a STEM graduate conditional on the classifiable keywords $\mathscr{K}_j^C$ collected from j's online job advert, and $\boldsymbol{X}_j$ includes controls, e.g. the pay frequency (daily, weekly, monthly...), the salary type (base pay, commission, bonus...), the month and year of the posting, whether the job is located in London, etc.

As shown in Table 1, the wage is posted explicitly in 61% of all job ads. However, introducing controls dramatically reduces the sample size, since, for instance, only 17% and 12% of the postings have minimum education and experience requirements, 46% have industry identifiers, etc. Hence, we present three sets of results: one obtained on a sample of almost 20 million vacancies, where we only require the vacancies to possess wage and four-digit UK SOC occupation identifiers in addition to some basic controls (Table 12). The second set of results, presented in Table 13, uses a much smaller sample of 222,451 postings in which we also observe the one/two-digit industry identifier, the precise county, and minimum education and experience requirements, and such that each occupation/industry combination has at least 2 observations. In the final set (Table 14), we do not require the industry identifier but require all the other controls already mentioned as well as the employer's name. This time, we ensure that each occupation/employer cell has at least 2 observations, which results in a sample of 62,511 observations.[28]For each year, we also drop the postings with the 1% lowest wages to remove outliers.

The first column of Table 12 is a plain regression of log hourly wages on the STEM job dummy with no controls. It suggests that, unconditionally, STEM jobs are associated with

[28]Requesting both the employer's name and the industry identifier, and ensuring that each unique occupation/industry/employer combination has at least two observations leads to a very small and unrepresentative sample dominated by a few large employers, like the NHS.

28% higher wages. Remember that we define STEM jobs as those whose recruiters look for STEM educated candidates with a higher probability than for non-STEM educated ones:

$$STEM_j = \mathcal{I}(\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) > \widetilde{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C))$$

Hence, not all STEM jobs are such that the recruiters seek STEM graduates with a 100% probability. A more flexible approach is therefore to use the probability of looking for a STEM graduate instead of the discrete STEM job indicator, i.e. the specification in eq.4.2 instead of eq.4.1. As shown in column (2), the premium offered for seeking a STEM graduate relative to a non-STEM one sharpens: a 10% pts. rise in the probability of looking for a STEM graduate is associated with a 3% pts. rise in the wage, so that as we go from looking for a non-STEM educated worker to seeking a STEM educated one, the wage offered rises by 32%. Note that this latter specification with the continuous probability instead of the discrete indicator also seems to provide a better description of the labour market dynamics since the R^2 rises from 5.5% to 5.9%.

The next column contrasts these results to the unconditional wage premium associated with working in a STEM occupation: 29%.

Columns (4) to (6) replicate these three specifications but now introducing some basic controls: a dummy for whether the job is located in London, the number of keywords in the description and the job title, the month and year of the posting, the pay frequency and the salary type. All estimates drop in size but remain highly significant. One of the reasons is probably that, as indicated in the previous subsection, a substantial part of STEM jobs are located in London, where wages are higher anyway because of higher living costs. Hence part of what appears as the STEM premium in columns (1) to (3) is the London premium, which disappears once we introduce the London dummy.

Table 12: The wage premium for STEM: regressions with basic controls

	Dependent variable: log(wage)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
STEM job	0.279***			0.237***			0.206***		0.156***			
	(0.000)			(0.000)			(0.000)		(0.018)			
$\widetilde{\Pr}$(STEM $\mid \mathscr{K}_j^C$)		0.319***			0.274***			0.259***		0.219***		0.233***
		(0.000)			(0.000)			(0.000)		(0.025)		(0.026)
STEM occupation			0.293***			0.222***	0.169***	0.162***				
			(0.000)			(0.000)	(0.001)	(0.001)				
STEM job							-0.104***		-0.021			
*STEM occ.							(0.001)		(0.031)			
$\widetilde{\Pr}$(STEM $\mid \mathscr{K}_j^C$)								-0.132***		-0.047		−0.059
*STEM occ.								(0.001)		(0.039)		(0.038)
Biology/Biomedicine											0.024*	−0.035***
											(0.013)	(0.013)
Computer Sciences											0.095***	0.027***
											(0.013)	(0.009)
Engineering											0.060***	0.002
											(0.010)	(0.006)
Maths/Statistics											0.032***	0.021***
											(0.006)	(0.007)
Technology											−0.051***	−0.091***
											(0.009)	(0.010)
Physics/Chemistry											−0.054***	−0.096***
											(0.015)	(0.014)
London				0.278***	0.279***	0.271***	0.276***	0.277***	0.219***	0.220***	0.215***	0.218***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.007)	(0.007)	(0.007)
No. Keywords				0.010***	0.010***	0.009***	0.009***	0.009***	0.004***	0.004***	0.002***	0.004***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Occupation dum.	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Year dum.	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month dum.	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pay Frequency	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Salary Type	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clustered s.e.	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Observations							19,856,575					
R^2	0.055	0.059	0.053	0.239	0.243	0.230	0.244	0.246	0.441	0.443	0.438	0.445
Adjusted R^2	0.055	0.059	0.053	0.239	0.243	0.230	0.244	0.246	0.441	0.443	0.438	0.445

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors in parentheses, clustered at the four-digit occupation level in columns (9) to (12). The wage is the average of the minimum and maximum hourly salaries posted. STEM job is a dummy for whether the job is classified as STEM. STEM occ. is a dummy for whether the job belongs to a STEM occupation. Regressions (1)-(3) include constants. Four-digit UK SOC occupations used. No. Keywords is the number of classified keywords collected from the job description and the job title.

Columns (7) and (8) consider specifications where we include together the STEM job indicator or $\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$, the STEM occupation dummy and an interaction between them. Note that we are still not controlling for a full set of four-digit UK SOC occupations. The results seem to indicate that there is a difference between the STEM premium offered in STEM and non-STEM occupations. For instance, column (8) suggests that the recruiter looking for a STEM graduate in a non-STEM occupation offers a 25.9% wage premium, whereas in a STEM occupation, he would offer a 16.2% wage premium for the fact that this is a STEM occupation and an additional 12.7% premium if looking for a STEM graduate, i.e. a 28.9% wage premium overall for a STEM job in a STEM occupation. However, as we introduce a full set of 368 four-digit UK SOC occupation dummies in columns (9) and (10), the interaction term becomes insignificant suggesting that the premium for STEM in STEM occupations is not statistically significantly different from the one in non-STEM occupations once we account for occupation fixed effects (note that standard errors in columns (9) and (10) are also clustered at the occupation level).

We continue by investigating whether different STEM domains command distinct premia in columns (11) and (12). This is an interesting question in itself which has already been investigated from the labour supply side in numerous papers. For instance, Greenwood et al. [30], who analyse the Labour Force Survey between March 2004 and December 2010, find that many qualifications have a higher labour market value if they are in a STEM subject. However, this general finding conceals an important amount of heterogeneity in returns to different STEM domains at different NQF levels. The authors conclude that "it is not enough to urge young people to study STEM subjects: they also need to understand that some STEM qualifications are more valuable than others." Other interesting contributions include, for instance, Webber [48] who looks at how average earnings vary by discipline in the US. Bratti et al. [11] use a British cohort study from 1970 to estimate wage returns by major studied. Gabe [27] takes a different approach. Instead of the discipline studied, he combines worker knowledge requirements from the O*NET with wage and demographic information from the U.S. Census American Community Survey. Although the results from all these papers are not directly comparable because of different data and methods, a general finding seems to be that sciences, especially Biology, Physics and Chemistry, are typically associated with lower earnings than Computer Sciences and Engineering.

In our case, we investigate the heterogeneity in STEM wage premia by defining Biology/Biomedicine, Computer Sciences, Engineering etc. indicators which are just equal to 1 if the vacancy description contains keywords belonging to the respective clusters (cf. Algorithm 1 for how keywords are classified into different STEM clusters). Column (11) includes occupation fixed effects and basic controls, but excludes the probability of looking for a STEM graduate. Technology and Physics/Chemistry seem to be associated with negative

wage premia, while the rest of STEM disciplines command positive ones. However, introducing the STEM probability and its interaction with the STEM occupation dummy in column (12) attenuates all coefficients and turns the one on Biology/Biomedicine negative. Further research could perhaps investigate heterogeneity in STEM wage premia in more details, however it is also important to remember that the separation of keywords into different clusters is imperfect and the results presented here are therefore only indicative.

In Tables 13 and 14, we decided to concentrate on the continuous measure of STEM requirements; the results with the discrete STEM job indicator are similar and available on request. We start by reproducing the analogues of columns (2), (3) and (10) from Table 12 to show what these specifications give on these much smaller and less representative samples. Columns (4) correspond to a regression that only includes full controls: education and experience requirements (in minimum years), a full set of counties instead of the London dummy, four-digit UK SOC occupations, and either one/two-digit industry identifiers in Table 13 or 6054 unique employers in Table 14. Columns (5) add the STEM probability and its interaction with the STEM occupation indicator terms. Finally, the specification in columns (6) also contains the different STEM domain dummies.

The main purpose of these sets of results is to show that the wage premium for STEM does not disappear even after introducing detailed controls for many other observable characteristics that affect wages. It certainly drops in magnitude as the influence of all these other factors is taken into account, but remains highly significant. The interaction term also remains insignificant. Most of the coefficients on the STEM domains in columns (6) go in the same direction as before, even though statistical significance drops, especially in the regression with employer fixed effects.

It is important to remember that all the results presented in this section are not causal as there could be an unobserved omitted variable - an analogue of the "ability" bias on the demand side, that is correlated with both wages and the probability of looking for a STEM graduate and is confounding our estimates even conditional all the controls introduced in Tables 13 and 14.[29]

Nevertheless, this section does provide evidence that controlling for detailed occupations, industries, employers, geographical locations, education and experience requirements, STEM jobs are still associated with higher wages in both STEM and non-STEM occupations, and that, conditional on occupation fixed effects, the premium for STEM does not differ depending on whether the occupation is STEM or non-STEM.

[29]It does not seem very plausible though that a recruiter would simply post, say, *"C++"* in his job advert, just because he thinks that a candidate who knows how to code in C++ is more able than one who does not, and not because the job genuinely requires knowledge of C++ or some other equivalent software that someone with knowledge of C++ could certainly easily learn.

Table 13: The wage premium for STEM: regressions with industry controls

	Dependent variable: log(wage)					
	(1)	(2)	(3)	(4)	(5)	(6)
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$	0.236***		0.187***		0.125***	0.129***
	(0.003)		(0.020)		(0.017)	(0.019)
STEM occ.		0.167***				
		(0.002)				
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$			−0.050		−0.037	−0.036
*STEM occ.			(0.033)		(0.027)	(0.026)
Biology/Biomedicine						−0.018
						(0.018)
Computer Sciences						0.0002
						(0.008)
Engineering						0.016**
						(0.006)
Maths/Statistics						0.018**
						(0.008)
Technology						−0.029***
						(0.010)
Physics/Chemistry						−0.045***
						(0.013)
London			0.203***			
			(0.011)			
No. Keywords			0.002***	0.001***	0.001***	0.001***
			(0.0003)	(0.0003)	(0.0003)	(0.0003)
Education				0.049***	0.049***	0.049***
				(0.003)	(0.003)	(0.003)
Experience				0.031***	0.030***	0.030***
				(0.003)	(0.003)	(0.003)
Occupation dum.	No	No	Yes	Yes	Yes	Yes
Industry dum.	No	No	No	Yes	Yes	Yes
County dum.	No	No	No	Yes	Yes	Yes
Year dum.	No	No	Yes	Yes	Yes	Yes
Month dum.	No	No	Yes	Yes	Yes	Yes
Pay Frequency	No	No	Yes	Yes	Yes	Yes
Salary Type	No	No	Yes	Yes	Yes	Ye
Clustered s.e.	No	No	Yes	Yes	Yes	Yes
Observations				222,451		
R^2	0.038	0.020	0.427	0.496	0.498	0.499
Adjusted R^2	0.038	0.020	0.426	0.494	0.497	0.497

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors in parentheses, clustered at the occupation. The wage is the average of the minimum and maximum hourly salaries posted. Education & experience requirements are in years (minimum required). Four-digit UK SOC occupations and one/two-digit SIC industries used.

Table 14: The wage premium for STEM: regressions with employer controls

	Dependent variable: log(wage)					
	(1)	(2)	(3)	(4)	(5)	(6)
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$	0.306***		0.172***		0.037***	0.039***
	(0.005)		(0.033)		(0.014)	(0.015)
STEM occ.		0.222***				
		(0.005)				
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$			0.0003		−0.039	−0.037
*STEM occ.			(0.052)		(0.026)	(0.026)
Biology/Biomedicine						−0.015
						(0.013)
Computer Sciences						−0.012
						(0.012)
Engineering						0.014
						(0.009)
Maths/Statistics						0.013
						(0.014)
Technology						−0.013
						(0.010)
Physics/Chemistry						−0.008
						(0.021)
London			0.202***			
			(0.015)			
No. Keywords			0.005***	0.003***	0.003***	0.003***
			(0.001)	(0.001)	(0.001)	(0.001)
Education				0.043***	0.043***	0.043***
				(0.003)	(0.003)	(0.003)
Experience				0.036***	0.036***	0.036***
				(0.002)	(0.002)	(0.002)
Occupation dum.	No	No	Yes	Yes	Yes	Yes
Employer dum.	No	No	No	Yes	Yes	Yes
County dum.	No	No	No	Yes	Yes	Yes
Year dum.	No	No	Yes	Yes	Yes	Yes
Month dum.	No	No	Yes	Yes	Yes	Yes
Pay Frequency	No	No	Yes	Yes	Yes	Yes
Salary Type	No	No	Yes	Yes	Yes	Ye
Clustered s.e.	No	No	Yes	Yes	Yes	Yes
Observations				62,511		
R^2	0.062	0.035	0.476	0.749	0.749	0.749
Adjusted R^2	0.062	0.035	0.473	0.719	0.719	0.719

*p<0.1; **p<0.05; ***p<0.01

Notes: Standard errors in parentheses, clustered at the employer level (6054 unique employers). The wage is the average of the minimum and maximum hourly salaries posted. Education & experience requirements are in years (minimum required). Four-digit UK SOC occupations.

As discussed in the introduction, previous studies often find that "STEM graduates [...] earn more than non-STEM graduates - but only if they work in science or finance occupations" (DIUS [17]). This finding is based on looking at the wages earned by STEM graduates without distinguishing between those among them who take up STEM jobs and those who end up in non-STEM ones. When looking from the labour supply side without making this important distinction, the wage premium for STEM that exists within non-STEM occupations could therefore be obscured since nothing prevents STEM graduates to take up non-STEM jobs, for which non-STEM graduates are also perfectly qualified and for which they therefore receive no premium. And, actually, 85% of all jobs within non-STEM occupations are non-STEM and therefore do not offer any wage premium for STEM skills even if they end up being filled with STEM graduates.

Hence, our results do not directly contradict, but rather extend previous findings. They are important because they suggest that STEM skills are valued and continue to contribute positively to productivity even within non-STEM occupations. Moreover, on the basis of conventional supply and demand, our results seem to be consistent with a shortage of STEM knowledge and skills across the economy and not only in STEM occupations.

4.4. **The STEM requirements of "Non-STEM" jobs.** We close this section by painting in more details the profile of STEM jobs belonging to non-STEM occupations which constitute the main object of interest in this paper.

In particular, we start by examining the top STEM requirements of STEM jobs belonging to:

- Chartered architectural technologists: *"Mechanical Engineering", "Engineering Management", "Civil Engineering", "Auto CAD", "Computer Aided Draughting/Design (CAD)", "Machinery", "HVAC", "Electrical! Engineering", "Engineering Design", "Revit", "Concept Development", "Technical Support", "Engineering consultation", "Systems Engineering", "Preventive Maintenance", "Mechanical Design", "Product Development", "Engineering Projects", "Engineering Support", "Lean Methods", "Process Design", "Manufacturing Industry Experience"...*
- Product, clothing and related designers: *"Computer Aided Draughting/Design(CAD)", "Concept Development", "Auto CAD", "Package Design", "Process Design", "Digital Design", "Product Development", "Product Design", "Concept Design and Development", "JavaScript", "User Interface (UI) Design", "Materials Design", "Java", "Prototyping", "Design Software", "Information Technology Industry Experience", "Revit", "Technical Drawings", "SQL", "Instrument Design", "CAD Design", "Set Design" ...*
- Management consultants and business analysts: *"SQL", "SAS", "Information Technology Industry Experience", "Data Warehousing", "Unified Modelling Language (UML)",*

"Scrum", "SQL Server", "Systems Analysis", "Data Modelling", "Extraction Transformation and Loading (ETL)", "Visual Basic", "SQL Server Reporting Services (SSRS)", "Validation", "Optimisation", "Systems Development Life Cycle (SDLC)", "Java", "Data Mining", "Process Design", "Agile Development", "Transact-SQL", "Extensible Markup Language (XML)", "Product Development", "Statistics", "Microsoft C#", "Relational Databases", "Big Data"...

- Graphic designers: *"Digital Design", "Concept Development", "Computer Aided Draughting/Design (CAD)", "Materials Design", "JavaScript", "AutoCAD", "HTML5", "Process Design", "User Interface (UI) Design" "Concept Design and Development", "Web Site Development", "jQuery", "Package Design", "Design Software", "Product Design", "Product Development", "Computer Software Industry Experience", "Technical Support", "Interface Design", "Prototyping", "Set Design", "Hypertext Preprocessor (php)", "3D Design", "3D Modelling", "Web Application Development"...*
- Actuaries, economists and statisticians: *"Statistics", "SAS", "Biostatistics", "SQL", "VisualBasic", "Bioinformatics", "Validation", "R", "Epidemiology", "Python", "C++", "Product Development", "Biology", "Optimisation", "PERL", "MATLAB", "Physics", "Mathematical Modelling", "Technical Support", "Pharmaceutical Industry Background", "Java", "Genomics", "Genetics", "UNIX", "Calibration", "LINUX", "Data Mining", "Model Building", "Experimental Design", "SIMULATION", "Predictive Models", "Relational Databases", "Experiments", "MySQL"...*
- Artists: *"Concept Development", "JavaScript", "Game Development", "Computer Aided Draughting/Design (CAD)", "Python", "Auto CAD", "3D Modelling", "Digital Design", "User Interface (UI) Design", "3D Design", "Autodesk", "Optimisation", "C++", "Microsoft C#", "3D Animation", "Technical Support", "Computer Software Industry Experience", "Troubleshooting", "Process Design", "Concept Design and Development", "Game Design", "ActionScript", "Materials Design", "Prototyping"...*

It seems that despite the fact that these recruiters are looking for STEM graduates with a higher probability than for non-STEM ones, many of the STEM skills and knowledge they require could actually be acquired with less training than a full-time STEM degree, and could therefore be taught to non-STEM graduates in order to make them suitable candidates for such positions.

Moreover, another interesting feature that distinguishes STEM jobs in STEM vs. non-STEM occupations is the percentage of keywords in the job description that are STEM. In STEM occupations, 60% of all keywords posted in a median STEM job advert are STEM, while in non-STEM occupations, this number is only 30% (means are 59.38% and 35.29% respectively).

Hence it seems that STEM recruiters within non-STEM occupations actually seek to combine STEM and non-STEM knowledge and skills in a certain combination that lies in between the STEM-dominated combination required in STEM occupations and the predominantly

non-STEM one asked for in non-STEM jobs (cf. our discussion of "hybrid" jobs in the Introduction).

5. Implications for STEM Skills & Knowledge Shortages

The previous section documents that a significant proportion of "non-STEM" employers may specifically look for STEM graduates not because they simply value their "foundation competencies", "logical approach to solving problems" or believe that STEM graduates are intrinsically more capable, but because a STEM education has equipped them with the skills and knowledge needed to write programs in C++ and JavaScript, create digital designs, develop user interfaces, work with Big Data, perform statistical analysis in SAS ... The jobs these employers advertise require and value STEM knowledge and skills despite being classified into "non-STEM" occupations. In reality, however, many of these STEM skills and knowledge could be learned with training that is less advanced than a full-time STEM degree and these "non-STEM" STEM recruiters actually want to combine them with non-STEM knowledge & skills.

In this section, we develop an abstract framework to think about the implications of these findings for higher education policies aimed at reducing STEM shortages. In particular, we illustrate how the STEM shortages experienced by "non-STEM" employers with STEM requirements and those that persist in traditional STEM occupations are related, and how teaching more STEM in non-STEM disciplines could help alleviate both.

5.1. **The Geometry of Skills & Knowledge Shortages.**

The first step in analyzing skills and knowledge shortages is to define them.

Unfortunately, no clear and objective definition exists in the academic literature where shortages are often understood as a phenomenon that "causes vacancies to remain open longer" (Haskel & Martin [36]). Unfilled vacancies constitute "dynamic shortages" which only persist until wages have risen such as to make enough people acquire the scarce skills and bring the labour market into equilibrium once again (Arrow and Capron [2]).

However, hiring difficulties, unfilled vacancies, wage rises, etc. are all potential *consequences* of shortages, not their proper *definition*. Hiring difficulties and unfilled vacancies may occur for reasons unrelated to shortages, like inefficient human resource recruiters, improper advertising of the job, etc., while raising wages is only *one of many* responses to shortages. For instance, the 2016/2017 Talent shortage survey conducted by ManpowerGroup [40] indicates that only 26% of employers respond to shortages by "paying higher salary packages to recruits". At the same time, 53% decide to "offer training and development to existing staff", 36% "recruit outside the talent pool", 28% "explore alternative sourcing strategies", 19% completely "change existing work models", etc.

Indeed, in practice, there is a great deal of confusion about both the meaning of shortages and the reactions to them on both sides of the labour market.

Green et al. (1998) [29] analyse the Employer Manpower and Skills Practices Survey (EMSPS) where employers were asked separate questions about experiencing (a) skills shortages, (b) difficulties in filling vacancies, and (c) deficiencies in the 'qualities' of their existing workforce. They find only a partial overlap in the responses to these questions, concluding that "to equate 'skill shortage' with 'hard-to-fill vacancy' may be a very risky assumption which, if falsely made, could lead to unsafe conclusions".

On the labour supply side, interviews and surveys of STEM students and graduates, analysed in BIS [42], reveal that most of them "start university with few career ideas". They typically choose to study a STEM discipline because of personal interest, enjoyment and/or aptitude. In their sample, less than a quarter of STEM graduates chose their degrees for "improved job prospects" and most of those who originally had career purposes in mind when enrolling in a STEM discipline, did so to keep their career options open. When it comes to applying for jobs, expected pay is certainly an important factor, but not the main motivating force. STEM graduates look primarily for "interesting work".[30]

Overall, it therefore remains unclear whether or not the potentially equilibrating wage adjustment mechanism is being used by employers and/or actually translating into more people acquiring the scarce skills and knowledge. In what follows, we therefore completely set these mechanisms apart and start from a basic definition of what a shortage is.

According to the British Government's Training Agency [1], a shortage occurs "when there are not enough people available with the skills needed to do the jobs which need to be done". We shall now try to translate this definition into an abstract framework, then employ it to conceptualize our empirical findings and think about education policies that could help reduce STEM shortages experienced in STEM and non-STEM occupations.

5.1.1. *Vacancies & Job seekers.* Let $\mathcal{V}$ denote the set of vacancies (empty jobs). The skills & knowledge requirements of any vacancy $j \in \mathcal{V}$ have two components:

- an absolute amount ϕ_j^v
- a composition $\boldsymbol{\theta}_j^v$: if skills & knowledge are m-dimensional, the *composition* required by job j is the $m \times 1$ vector:

$$\boldsymbol{\theta}_j^v = (\theta_{j1}^v, \theta_{j2}^v, ..., \theta_{jm}^v)$$

such that $\sum_{l=1}^m \theta_{jl}^v = 1$ and $\theta_{jl}^v \in [0,1]\ \forall l$ is the proportion of j's overall requirements in the $l-$dimension.

[30]The academic literature also contains many contributions showing that financial incentives have little or no impact of student learning choices at all education levels, cf. Fryer [26] and references therein.

Let Ω be the m-dimensional skills & knowledge space. The location of vacancy j in Ω is determined by the vector $v_j = \phi_j^v \boldsymbol{\theta}_j^v$ with components $v_{jl} = \phi_j^v \theta_{jl}^v$.

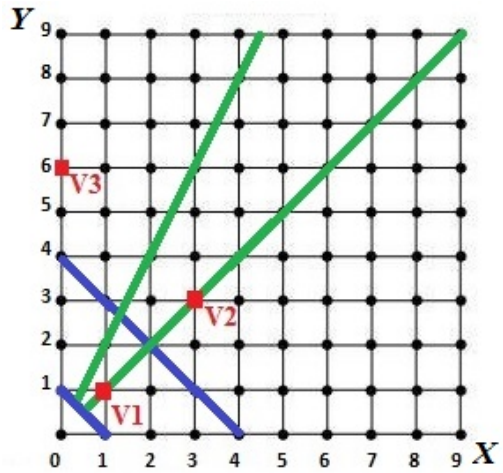
Figure 5.1 illustrates the idea on a two dimensional lattice.

Along each blue line (and we have only shown two for clarity), the same amount of skills & knowledge but a different composition are required. As we move from the left to the right, the composition is tilted towards the X dimension because its loading on the latter increases, while the share allocated to the Y dimension decreases.

Along each green line, the same composition but a different amount are required. As we move towards the North-East, the amount of skills & knowledge required increases.

For example, vacancies V_2 and V_3 require the same amount of skills & knowledge $\phi_2^v = \phi_3^v = 6$ but different compositions. Vacancies V_1 and V_2 have the same compositions $\boldsymbol{\theta}_1^v = \boldsymbol{\theta}_2^v = (0.5, 0.5)$ but require different amounts.

Figure 5.1 Skills & Knowledge space



Notes: V2 and V3 require the same amount of skills & knowledge but different compositions. V1 and V2 ask for the same composition but different amounts.

Let $\mathcal{S}$ denote the set of job seekers. As with vacancies, the location of candidate i in the space Ω is characterized by $s_i = \phi_i^s \boldsymbol{\theta}_i^s$ where ϕ_i^s is the amount of skills & knowledge possessed and $\boldsymbol{\theta}_i^s$ the m-dimensional composition vector.

An employer requiring amount ϕ_j^v and composition $\boldsymbol{\theta}_j^v$ to fill vacancy j might be indifferent between a certain subset of candidates located in $Z_j \subset \Omega$, where Z_j could be influenced by many things but, for clarity, is assumed to only depend on the vacancy's location here, i.e. $Z_j \coloneqq Z(v_j)$.

Formally, let $\omega = \phi\boldsymbol{\theta}$ be a generic element of Ω (which we denote as v and s when referring to vacancies and graduates respectively).

Definition 1. The *qualified subset* for vacancy j, Z_j is such that for any two elements $\omega \neq \omega'$ with $\omega \in Z_j$ and $\omega' \in Z_j$:

$$\omega \sim_j \omega'$$

i.e. the recruiter for vacancy j is indifferent between the two in terms of knowledge and skills.

In practice, we could think of Z_j as the subset of candidates for vacancy j such that it is no longer differences in the knowledge & skills that these candidates possess which will be the main determinant of the hiring decision. Other worker characteristics such as work styles, personality will allow the recruiter to select the best fit for his vacancy. However, here we

abstract from all this and concentrate on skills & knowledge in terms of which the candidates all seem equally qualified to the employer.

The existence of such subsets implies that a candidate may be simultaneously qualified for several vacancies belonging to the same or different occupations. In this case, the question of establishing whether or not there are "enough" qualified people available to "do the jobs which need to be done", i.e. to fill all open vacancies, becomes non-trivial as we cannot simply count the numbers of vacancies and job seekers at every ω and declare a shortage if vacancies outnumber candidates.

To determine whether vacancies located at a specific point in the skills & knowledge space experience a shortage, we start by characterizing the measure space in which vacancies and job seekers coexist.

For simplicity, suppose that Ω is discrete.

The distribution of the job seekers defines a measure P on Ω. For instance, if the pool of job seekers is such that none of them is located at ω, i.e. $s_i \neq \omega$ for all $i \in \mathcal{S}$, we have $P(\omega) = 0$. More generally:

$$P(\omega) = |\{i \in \mathcal{S} | s_i = \omega\}| \text{ and } P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = |\mathcal{S}| \tag{5.1}$$

where $|.|$ denotes the cardinality of a set.

In a similar way, we can define a measure Q for the distribution of the vacancies:

$$Q(\omega) = |\{j \in \mathcal{V} | v_j = \omega\}| \text{ and } Q(\Omega) = \sum_{\omega \in \Omega} Q(\omega) = |\mathcal{V}| \tag{5.2}$$

Definition 2. Vacancies located at $\Delta \in \Omega$ experience a shortage if:

$$Q(\Delta) + \sum_{\{\omega \in \mathcal{H}_\Delta\}} Q(\omega) > P(Z_\Delta) + \sum_{\{\omega \in \mathcal{L}_\Delta\}} P(\omega) \tag{5.3}$$

where $\mathcal{H}_\Delta \coloneqq \{\omega \in \Omega | P(Z_\omega \cap Z_\Delta) \neq 0, \omega \neq \Delta\}$ and $\mathcal{L}_\Delta \coloneqq \{\omega \in \Omega | \omega \in Z_u \text{ for } u \in \mathcal{H}_\Delta, \omega \notin Z_\Delta\}$. Note that since $Z_j \coloneqq Z(v_j)$, $Z_j = Z_h = Z_\Delta$ for any $j \neq h$ such that $v_j = v_h = \Delta$.

The left hand side of eq.5.3 gives the total demand for candidates qualified for vacancies located at Δ. The first term is simply the number of vacancies at Δ. The second one counts all the other vacancies that also want to hire job seekers qualified for vacancies at Δ. This subset of vacancies is denoted by $\mathcal{H}_\Delta$. On the right hand side, the first term gives the number of candidates qualified for vacancies at Δ, while the second one adjusts this number for the fact that vacancies in $\mathcal{H}_\Delta$, i.e. which compete with vacancies at Δ for the job seekers in Z_Δ, also have access to a pool of candidates that are qualified for them but unqualified for vacancies at Δ, and for which they do not compete with Δ-vacancies.

Example 3. To fully understand the condition for a shortage contained in eq.5.3, we can look at a simple example with a two-dimensional space in which the problem can be inspected visually.

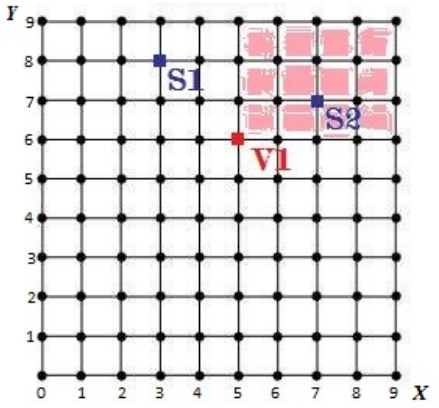
For clarity, let's also assume the following specific form for the qualified subsets, illustrated in Figure 5.2:

$$Z_j = \{\omega \in \Omega | \omega_l \geqslant v_{jl}, \forall l = 1, ..., n\} \tag{5.4}$$

Intuitively, eq.5.4 corresponds to the subset of candidates who have at least as much skills & knowledge in each dimension as what the vacancy requires.

In Figure 5.2, vacancy 1 is located at $(5, 6)$, while the job seekers are at $s_1 = (3, 8)$ and $s_2 = (7, 7)$. The shaded area to the North-East of vacancy 1 corresponds to Z_1 as defined in eq.5.4. Only candidate 2 belongs to Z_1. Candidate 1 is unqualified because $s_{11} < v_{11}$. In particular, s_1 has the right amount of skills & knowledge ($\phi_1^s = \phi_1^v = 11$) but the wrong composition: $\boldsymbol{\theta}_1^s = (3/11, 8/11)$ versus $\boldsymbol{\theta}_1^v = (5/11, 6/11)$. The skills & knowledge composition of s_2 assigns equal weights to both dimensions. Although the composition required by the vacancy is slightly tilted towards the vertical dimension compared to the one possessed by s_2, he is still qualified for the job according to eq.5.4 because he has more overall skills & knowledge and is located such that $s_{21} > v_{11}$ and $s_{22} > v_{11}$.

Figure 5.2 Qualified subset

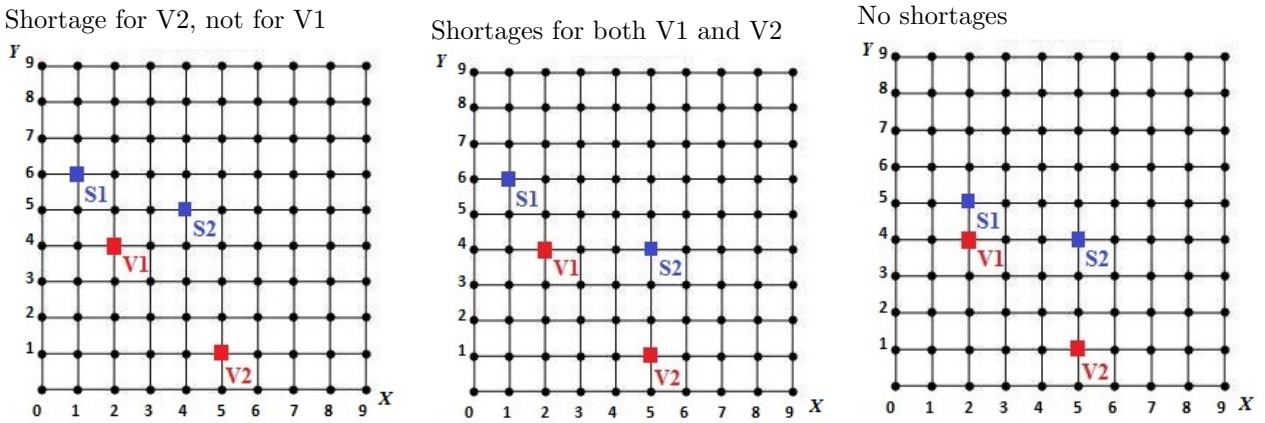


Notes: The shaded area marks the subset of candidates who would be qualified for vacancy V1 according to eq. 5.4

Example 4. Figure 5.3 illustrates how the condition for determining the presence of a shortage in eq. 5.3 works in this simple abstract setting by presenting three possible scenarios with two vacancies and two job seekers. Let the L, M, and R - subscripts stand for its left, middle, and right panels.

In the left wing panel we have $P^L(\omega) = 1$ for $\omega = \{s_1, s_2\}$ with $s_1 = (1, 6)$ and $s_2 = (4, 5)$ and $P^L(\omega) = 0$ for any other $\omega \in \Omega$. For the vacancies, $Q^L(\omega) = 1$ for $\omega = \{v_1, v_2\}$ with $v_1 = (2, 4)$ and $v_2 = (5, 1)$ and $Q^L(\omega) = 0$ for $\omega \neq \{v_1, v_2\}$. Note that vacancies require exactly the same amount of skills & knowledge ($\phi_1^v = \phi_2^v = 6$) but different compositions: $\boldsymbol{\theta}_1^v = (1/3, 2/3)$ and $\boldsymbol{\theta}_2^v = (5/6, 1/6)$. Furthermore, the qualified subsets are such that $P^L(Z_1) = 1$ and $P^L(Z_2) = 0$. Vacancy 2 experiences a shortage since both potential candidates have inadequate compositions despite having more skills & knowledge that what v_2 requires: $\phi_1^s = 7$ and $\phi_2^s = 9$. Since $P^L(Z_1 \cap Z_2) = 0$, eq.5.3 reads: $1 + 0 > 0 + 0$ at $\Delta = v_2$ signaling a shortage for vacancy 2. By contrast, vacancy 1 does not experience a shortage; the qualified candidate is s_2, and there is *enough* of him because he is not also qualified for vacancy 2.

Figure 5.3 Illustrating the condition for shortages in eq. 5.3



Notes: Three possible scenarios with two workers and two vacancies in a two-dimensional skills & knowledge space are illustrated. As S2 becomes qualified for both vacancies (going from the left to the middle panel), there is no longer "enough" of him, so that both V1 and V2 experience shortages. Moving from the middle to the right panel, both shortages are eliminated by simply making S1 qualified for V1, so that there are enough qualified workers at the level of the economy to simultaneously fill both vacancies.

Eq.5.3 in this case gives $1 + 0 \leq 1 + 0$ since $Q^L(v_1) = P^L(Z_1) = P^L(s_2) = 1$ and the second terms on both sides are still equal to 0 because $P^L(Z_1 \cap Z_2) = 0$.

In the middle panel, we simply change the location of s_2 from $(4, 5)$ to $(5, 4)$, i.e. keeping $\phi_2^s = 9$ but slightly changing his skills & knowledge composition. Candidate 2 is now the *only* qualified candidate for both vacancies and so there is no longer *enough* of him. Indeed, now $P^M(Z_1 \cap Z_2) = 1$ and eq.5.3 becomes $1 + 1 > 1 + 0$ for both vacancies, indicating shortages.

Finally, in the right wing of fig 5.3 we move job seeker 1 from $(1, 6)$ to $(2, 5)$, keeping everything else as in the middle panel. s_1 is now in Z_1, i.e. qualified for vacancy 1, while still remaining outside Z_2. Graphically, it is obvious that there are no shortages because there are *enough* qualified candidates to fill both vacancies simultaneously. Simply assign s_1 to v_1 and s_2 to v_2. The condition for a shortage in eq.5.3 is violated for both vacancies. For vacancy 2, the equation reads $1 + 1 \leq 1 + 1$ since $Q^R(v_2) = 1$, $P^R(Z_2) = 1$, and $P^R(Z_1 \cap Z_2) = 1$. It is important not to forget the right hand side adjustment $P^R(s_1) = 1$. Indeed, although s_2 seems to be over-demanded since he is qualified for both vacancies so that total demand for him is $Q^R(v_1) + Q^R(v_2) = 2$, it would be wrong to conclude that v_2 experiences a shortage because s_2 is the only qualified applicant for it. The reason is that, contrary to the situation depicted in the middle panel, vacancy 1 now has an alternative qualified candidate: s_1.

5.2. **Implications for Higher Education provision policies.** We can think about s_2 as the STEM graduate and s_1 as the non-STEM graduate, X as the STEM dimension and Y as the non-STEM dimension, v_2 as the traditional STEM occupation STEM job, and of v_1 as a "non-STEM" STEM job.

The above example illustrates two points:

- the existence of a shortage cannot be established by looking at some vacancies and job seekers in isolation, it has to take into account their interdependence at the level of the economy. Hence, in order to understand why STEM shortages arise and propose adequate policies to eliminate them, we need to go beyond STEM graduates and STEM occupations, and include non-STEM graduates and non-STEM occupations.
- shortages can be solved by changing the location of "not-in-shortage" graduates: in the middle panel of fig. 5.3, the skills & knowledge composition of s_1 allocates too little to the STEM dimension for him to be qualified even for v_1- the "non-STEM" job that nevertheless requires a certain amount of STEM knowledge and skills. Hence both v_1 and v_2 have to compete for s_2. In the right panel, we simply change the composition of s_1 from (1/7, 6/7) to (2/7, 5/7) without adding any skills & knowledge. This solves shortages for both v_1 and v_2 because they no longer have to compete for the STEM graduate s_2, and also gives a job to the non-STEM graduate.

These points imply that STEM shortages are not only about "not enough" STEM graduates, but also about "not enough" STEM skills & knowledge taught in non-STEM disciplines.

A key implication is that the solution to STEM shortages is not only and necessarily to encourage more students to enroll into STEM degrees, which many of them will avoid, however high the rewards may be, because following advanced STEM classes for several years of their lives might be too difficult and/or uninteresting. Instead, introducing more mandatory, or at least optional, STEM modules into non-STEM disciplines could help alleviate shortages by allowing students to enroll in non-STEM degrees while still graduating with the employer-desired amount of STEM knowledge and skills. Furthermore, this policy could help alleviate shortages in traditional STEM occupations, since if there are more non-STEM graduates with appropriate STEM training, "non-STEM" STEM employers will be less likely to look specifically for STEM graduates, who may therefore have to seek jobs in the traditional STEM occupations more often.

6. Conclusion & Future Research

This paper aims to contribute to the debate on whether the fact that, in the UK, less than half of STEM graduates work in non-STEM occupations should be considered as a problem or not necessarily so, and, if yes, what type of education provision policy initiatives could help resolve it.

We develop a new approach to identifying STEM jobs through the keywords collected from online vacancy descriptions, and not, as is typically done, by classifying occupations discretely into STEM vs. non-STEM, then considering all the jobs belonging to the first group as "STEM" and the rest as "non-STEM". This approach is made possible by having

access to a large dataset, collected by the firm *Burning Glass Technologies*, which contains information on all vacancies posted online in the UK between 2012 and 2016.

Our job level analysis shows that it is wrong to equate STEM jobs with STEM occupations: 35% of all STEM jobs belong to non-STEM occupations. Moreover, this leads to underestimating the overall demand for STEM knowledge and skills since STEM jobs outnumber jobs in STEM occupations, e.g. by half a million STEM employment opportunities in 2015. We also find that when seeking STEM graduates, recruiters in non-STEM occupations offer to pay higher wages and, conditional on occupation fixed effects, this premium is not statistically significantly different from the one offered for STEM knowledge and skills within STEM occupations.

Although, these findings suggest that the leakage from the STEM pipeline may be less problematic than typically thought because around 15% of all recruiters in non-STEM occupations do require and value STEM knowledge and skills, the issue remains problematic for two main reasons.

Firstly, nothing prevents STEM educated job seekers to take up non-STEM jobs within non-STEM occupations, for which non-STEM graduates are also qualified and no STEM premium is offered.

Secondly, we find that the STEM skills and knowledge posted in STEM vacancies within non-STEM occupations go beyond *"Problem Solving", "Analytical Skills"...* but, in many cases, could be acquired with less training than a full time STEM degree. Moreover, STEM recruiters within non-STEM occupations actually wish to combine STEM knowledge and skills with non-STEM ones to a larger extent that their counterparts in STEM occupations. Hence, a more efficient way of satisfying STEM demand within non-STEM occupations could be to teach more STEM in non-STEM disciplines so as to make non-STEM graduates qualified for a set of jobs within non-STEM occupations for which they only lack the STEM skills while already possessing the required non-STEM ones. We construct an abstract framework to illustrate how this reform could reduce STEM shortages in both STEM and non-STEM occupations.

Although the main focus of this paper is on "high level" STEM jobs – jobs that belong to managerial, professional and associate professional occupations for which a university degree is typically required, our analysis indicates that 25% of all STEM employment opportunities in the UK are not "high level". Examining the O*NET Knowledge scales, Rothwell [45] gets a similar result for the US, finding that 30% of STEM positions there are "blue-collar". He argues that "the excessively professional definition of STEM jobs has led to missed opportunities to identify and support valuable training and career development." This could also be the case in the UK. Hence, future research should spend more time investigating non-graduate STEM job openings in the UK as well.

Moreover, future research could also try to merge the analysis of STEM demand presented here with a similar analysis on the labour supply side, perhaps using a dataset like LinkedIn. Nowadays, many people acquire STEM human capital not through formal education but other channels like self-study, online courses, internships... This makes the assessment of the actual supply of STEM knowledge & skills more complex than simply counting the number of STEM graduates. Although, the existing STEM literature recognizes this as a problem, so far, no attempts seem to have been made to deal with it and "STEM skills and knowledge" continue to be used interchangeably with "STEM qualifications". Clearly, this has the same flaws as equating "STEM jobs" with "STEM occupations".

References

[1] Training Agency. 'Labour market and skill trends 1991/1992'. *Employment Department*, 1990.

[2] Kenneth J. Arrow and William M. Capron. 'Dynamic shortages and price rises: the Engineer-scientist'. *Quarterly Journal of Economics*, 73(2), 1959.

[3] General Assembly and Burning Glass Technologies. 'Blurring lines: How business and technology skills are merging to create high opportunity hybrid jobs'. 2015.

[4] R. H. Baayen. *'Analyzing Linguistic Data: A Practical Introduction to Statistics Using R'*. Cambridge University Press, 2008.

[5] Joshua Ballance, Alicia Sasser Modestino, and Daniel Shoag. 'Downskilling: Changes in employer skill requirements over the business cycle'. *Labour Economics*, 2016.

[6] Joshua Ballance, Alicia Sasser Modestino, and Daniel Shoag. 'Upskilling: Do employers demand greater skill when workers are plentiful?'. *Working Paper*, February 2016.

[7] Kenneth Benoit and Paul Nulty. *quanteda: Quantitative Analysis of Textual Data*, 2015. R package version 0.7.2-1.

[8] BIS. 'Research into the need for and capacity to deliver STEM related apprenticeship provision in England'. *BIS Research Paper Number 171*, March 2014.

[9] Derek Bosworth, Clare Lyonette, Rob Wilson, Marc Bayliss, and Simon Fathers. 'The supply of and demand for high-level STEM skills'. *UKCES Evidence Report 77*, November, 2013.

[10] Amber E. Boydstun, Timothy P. Jurka, Loren Collingwood, Emiliano Grossman, and Wouter van Atteveldt. *RTextTools: Automatic Text Classification via Supervised Learning*, 2014. R package version 1.4.2.

[11] Massimiliano Bratti, Robin Naylor, and Jeremy Smith. 'Heterogeneities in the returns to degrees: Evidence from the British cohort study 1970'. *Departmental Working Paper*, December 2008.

[12] Erik Brynjolfsson and Andrew McAfee. 'Big data: The management revolution'. *Harvard Business Review*, October, 2012.

[13] A.P. Carnevale, T. Jayasundera, and D. Repnikov. 'Understanding online job ads data: a technical report'. *George Town University*, 2014.

[14] Arnaud Chevalier. 'To be or not to be... a scientist?'. *IZA Discussion Paper No. 6353*, February 2012.

[15] Yu-Wei Chiu. *'Machine Learning with R Cookbook'*. Packt Publishing, March, 2015.

[16] David Deming and Lisa B. Kahn. 'Firm heterogeneity in skill demands'. June 2016.

[17] Universities Department for Innovation and Skills. 'The demand for science, technology, engineering and mathematics (STEM) skills '. January 2009.

[18] Higher Education Funding Council For England. 'Guide to funding 2015-16: How HEFCE allocates its funds'. 2015.

[19] Ingo Feinerer, Kurt Hornik, and David Meyer. *tm: Text Mining Package*, 2015. R package version 0.6-2.

[20] Ingo Feinerer, Kurt Hornik, and David Meyer. 'Text mining infrastructure in R'. *Journal of Statistical Software*, 25(5), March 2008.

[21] Ronen Feldman and James Sanger. *'The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data'*. Cambridge University Press.

[22] Ian Fellows. *wordcloud: Word Clouds*, 2014. R package version 2.5.

[23] UK Commission for Employment and Skills. 'Reviewing the requirement for high level STEM skills'. *UKCES Evidence Report 94*, July, 2015.

[24] National Center for O*NET Development for USDOL. *The O*NET Content Model: Detailed Outline with Descriptions*.

[25] Carl Benedikt Frey and Michael A. Osborne. 'The future of employment: How susceptible are jobs to computerisation?'. *Technological Forecasting and Social Change*, 114, January 2017.

[26] Roland G. Jr. Fryer. 'Financial incentives and student achievement: evidence from randomized trials'. *Quarterly Journal of Economics*, 126(4), 2011.

[27] Todd M. Gabe. 'Knowledge and earnings'. *Journal of Regional Science*, 49(3), 2009.

[28] Marek Gagolewski and Bartek Tartanus. *R package stringi: Character string processing facilities*, 2016.

[29] F. Green, S. Machin, and D. Wilkinson. 'The meaning and determinants of skills shortages'. *Oxford Bulletin of Economics and Statistics*, 60, 2, 1998.

[30] Charley Greenwood, Matthew Harrison, and Anna Vignoles. 'The labour market value of STEM qualifications and occupations'. *Royal Academy of Engineering, Department of Quantitative Social Science, Institute of Education*, 2011.

[31] J.A. Hartigan and M. A. Wong. 'Algorithm as 136: A K-means clustering algorithm'. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1979.

[32] Brad J. Hershbein and Lisa B. Kahn. 'Do recessions accelerate routine-biased technological change? Evidence from vacancy postings'. *NBER Working Paper 22762*, October 2016.

[33] Kurt Hornik. *NLP: Natural Language Processing Infrastructure*, 2015. R package version 0.1-8.

[34] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *'An Introduction to Statistical Learning'*. Springer Texts in Statistics, 2013.

[35] Friedman Jerome, Trevor Hastie, and Robert Tibshirani. *'The Elements of Statistical Learning, Data Mining, Inference, and Prediction'*. Springer Series in Statistics, 2008.

[36] J.Haskel and C. Martin. 'Do skill shortages reduce productivity? Theory and evidence from the United Kingdom'. *The Economic Journal*, 103, 1993.

[37] Roger Koenker and Pin Ng. *SparseM: Sparse Linear Algebra*, 2015. R package version 1.7.

[38] C. Levy and L. Hopkins. 'Shaping up for innovation: are we delivering the right skills for the 2020 knowledge economy?'. *The Work Foundation. London*, 2010.

[39] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *'An Introduction to Information Retrieval'*. Cambridge University Press, 2009.

[40] ManpowerGroup. 2016/ 2017 Talent shortage survey. 2016.

[41] Geoff Mason. 'Science, Engineering and Technology Technicians in the UK Economy'. *National Institute of Economic and Social Research, London*, 2012.

[42] Robin Mellors-Bourne, Helen Connor, and Charles Jackson. 'STEM graduates in non STEM jobs'. *BIS Research Paper number 30*, March, 2011.

[43] Andrew Reamer. 'Using real-time labor market information on a nationwide scale'. *Jobs For the Future*, April, 2013.

[44] Jonathan Rothwell. 'Still searching: Job vacancies and STEM skills'. *Brookings*, July, 2014.

[45] Jonathan Rothwell. 'The hidden STEM economy'. *Brookings*, June, 2013.

[46] Gaston Sanchez. *'Handling and Processing Strings in R'*. Trowchez Editions. Berkeley, 2013.

[47] Matthew Sigelman. 'Why the STEM gap is bigger than you think'. April, 2016.

[48] Douglas Webber. 'Is the return to education the same for everybody?'. *IZA World of Labor 2014: 92*, October 2014.

[49] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2015. R package version 1.0.0.

7. Appendix

Table 15: Occupational distribution of STEM jobs in 2015

Code	**Name (4-digit UK SOC)**	*STEM occ.*	*STEM density*	*% STEM jobs*
1115	Chief executives and senior officials	0	15.1	0.07
1116	Elected officers and representatives	0	63.06	0.04
1121	Production managers and directors in manufacturing	1	60.06	1.41
1122	Production managers and directors in construction	0	78.56	1.09
1123	Production managers and directors in mining and energy	1	64.5	0.05
1131	Financial managers and directors	0	3.88	0.09
1132	Marketing and sales directors	0	16.61	0.54
1133	Purchasing managers and directors	0	15.83	0.13
1134	Advertising and public relations directors	0	2.06	0.01
1135	Human resource managers and directors	0	2.94	0.03
1136	Information technology and telecommunications directors	1	33.39	0.13
1139	Functional managers and directors n.e.c.	0	16.32	0.08
1150	Financial institution managers and directors	0	9.12	0.02
1161	Managers and directors in transport and distribution	0	25.19	0.08
1162	Managers and directors in storage and warehousing	0	29.86	0.14
1171	Officers in armed forces	0	13.81	0.02
1172	Senior police officers	0	19.4	0
1173	Senior officers in fire, ambulance, prison and related services	0	51.46	0.06
1181	Health services and public health managers and directors	0	9.24	0.08
1184	Social services managers and directors	0	1.24	0
1190	Managers and directors in retail and wholesale	0	10.57	0.27
1211	Managers and proprietors in agriculture and horticulture	0	5.82	0
1213	Managers and proprietors in forestry, fishing and related services	0	30.49	0
1221	Hotel and accommodation managers and proprietors	0	10.38	0.02
1223	Restaurant and catering establishment managers and proprietors	0	1.7	0.01

1224	Publicans and managers of licensed premises	0	3.21	0
1225	Leisure and sports managers	0	5.53	0.01
1226	Travel agency managers and proprietors	0	3.6	0
1241	Health care practice managers	0	0.04	0
1242	Residential, day and domiciliary care managers and proprietors	0	0.27	0
1251	Property, housing and estate managers	0	23.43	0.25
1252	Garage managers and proprietors	0	13.74	0
1253	Hairdressing and beauty salon managers and proprietors	0	0.75	0
1254	Shopkeepers and proprietors ¡U+0096¿ wholesale and retail	0	12.87	0.02
1255	Waste disposal and environmental services managers	1	47.04	0.01
1259	Managers and proprietors in other services n.e.c.	0	42.66	2.44
2111	Chemical scientists	1	93.39	0.2
2112	Biological scientists and biochemists	1	64.12	0.54
2113	Physical scientists	1	75.65	0.1
2114	Social and humanities scientists	0	7.62	0.01
2119	Natural and social science professionals n.e.c.	1	88.69	0.27
2121	Civil engineers	1	98.66	1.72
2122	Mechanical engineers	1	99.39	1.15
2123	Electrical engineers	1	99.66	1.15
2124	Electronics engineers	1	98.19	0.3
2126	Design and development engineers	1	99.11	2.5
2127	Production and process engineers	1	94.62	0.5
2129	Engineering professionals n.e.c.	1	80.37	0.97
2133	IT specialist managers	1	54.71	0.69
2134	IT project and programme managers	1	50.92	1.06
2135	IT business analysts, architects and systems designers	1	79.28	5.36
2136	Programmers and software development professionals	1	91.41	14.61
2137	Web design and development professionals	1	90.87	5.18
2139	Information technology and telecommunications professionals n.e.c.	1	77.7	3
2141	Conservation professionals	1	49.7	0.05
2142	Environment professionals	1	57.44	0.1
2150	Research and development managers	1	58.88	0.25
2211	Medical practitioners	0	5.42	0.15
2212	Psychologists	1	0.14	0
2213	Pharmacists	0	4.98	0.02
2214	Ophthalmic opticians	0	0.6	0
2215	Dental practitioners	0	0.31	0
2216	Veterinarians	1	0.3	0
2217	Medical radiographers	0	4.28	0.01
2218	Podiatrists	0	0.07	0
2219	Health professionals n.e.c.	0	6.75	0.03
2221	Physiotherapists	0	0.07	0

2222	Occupational therapists	0	0	0
2223	Speech and language therapists	0	0	0
2229	Therapy professionals n.e.c.	0	8.18	0.01
2231	Nurses	0	0.18	0.02
2232	Midwives	0	0.56	0
2311	Higher education teaching professionals	0	10.63	0.04
2312	Further education teaching professionals	0	6.21	0.04
2314	Secondary education teaching professionals	0	4.54	0.13
2315	Primary and nursery education teaching professionals	0	0.15	0
2316	Special needs education teaching professionals	0	0.54	0
2317	Senior professionals of educational establishments	0	2.99	0.02
2318	Education advisers and school inspectors	0	6.05	0.01
2319	Teaching and other educational professionals n.e.c.	0	3.01	0.05
2412	Barristers and judges	0	19.02	0.01
2413	Solicitors	0	2.8	0.07
2419	Legal professionals n.e.c.	0	2.85	0.04
2421	Chartered and certified accountants	0	0.35	0.01
2423	Management consultants and business analysts	0	25.33	1.23
2424	Business and financial project management professionals	0	13.84	0.22
2425	Actuaries, economists and statisticians	0	34.69	0.13
2426	Business and related research professionals	0	33.76	0.24
2429	Business, research and administrative professionals n.e.c.	0	46.84	0.43
2431	Architects	1	65.89	0.42
2432	Town planning officers	1	65.14	0.24
2433	Quantity surveyors	0	29.96	0.58
2434	Chartered surveyors	0	64.7	0.66
2435	Chartered architectural technologists	0	85.42	0.24
2436	Construction project managers and related professionals	0	56.19	0.14
2442	Social workers	0	0.15	0
2443	Probation officers	0	0.16	0
2444	Clergy	0	4.58	0
2449	Welfare professionals n.e.c.	0	2.51	0
2451	Librarians	0	8.89	0.01
2452	Archivists and curators	0	11.76	0.01
2461	Quality control and planning engineers	1	90.78	1.11
2462	Quality assurance and regulatory professionals	1	49.94	0.87
2463	Environmental health professionals	1	15	0
2471	Journalists, newspaper and periodical editors	0	20.48	0.14
2472	Public relations professionals	0	2.61	0.01
2473	Advertising accounts managers and creative directors	0	0.91	0
3111	Laboratory technicians	1	54.29	0.38
3112	Electrical and electronics technicians	1	97.63	0.11
3113	Engineering technicians	1	93.59	2.09
3114	Building and civil engineering technicians	1	93.89	0.2

3115	Quality assurance technicians	1	85.95	0.5
3116	Planning, process and production technicians	1	81.93	0.18
3119	Science, engineering and production technicians n.e.c.	1	75.11	1.24
3121	Architectural and town planning technicians	1	53.66	0.15
3122	Draughtspersons	1	84.23	0.58
3131	IT operations technicians	1	72.12	2.74
3132	IT user support technicians	1	71.93	3.3
3213	Paramedics	0	1	0
3216	Dispensing opticians	0	0.35	0
3217	Pharmaceutical technicians	1	2.88	0
3218	Medical and dental technicians	1	34.27	0.16
3219	Health associate professionals n.e.c.	0	2.42	0.03
3231	Youth and community workers	0	0.69	0
3233	Child and early years officers	0	0.6	0
3234	Housing officers	0	2.4	0.01
3235	Counsellors	0	1.14	0
3239	Welfare and housing associate professionals n.e.c.	0	2.83	0.03
3311	NCOs and other ranks	0	9.7	0.02
3312	Police officers (sergeant and below)	0	28.89	0.02
3313	Fire service officers (watch manager and below)	0	61.55	0.01
3314	Prison service officers (below principal officer)	0	5.25	0
3315	Police community support officers	0	8.06	0
3319	Protective service associate professionals n.e.c.	0	41.16	0.07
3411	Artists	0	23.46	0.04
3412	Authors, writers and translators	0	12.13	0.15
3413	Actors, entertainers and presenters	0	11.36	0.08
3414	Dancers and choreographers	0	1.01	0
3415	Musicians	0	12.41	0.02
3416	Arts officers, producers and directors	0	10.58	0.05
3417	Photographers, audio-visual and broadcasting equipment operators	0	18.05	0.05
3421	Graphic designers	0	18.53	0.25
3422	Product, clothing and related designers	0	45.62	0.31
3441	Sports players	0	11.39	0.01
3442	Sports coaches, instructors and officials	0	5.63	0.02
3443	Fitness instructors	0	0.2	0
3511	Air traffic controllers	0	42.98	0
3512	Aircraft pilots and flight engineers	0	35.17	0.01
3513	Ship and hovercraft officers	0	26.74	0.02
3520	Legal associate professionals	0	1.52	0.02
3531	Estimators, valuers and assessors	0	45.21	0.97
3532	Brokers	0	14.29	0.05
3533	Insurance underwriters	0	14.56	0.06
3534	Finance and investment analysts and advisers	0	7.59	0.19

3535	Taxation experts	0	6.03	0.03
3536	Importers and exporters	0	11.58	0.01
3537	Financial and accounting technicians	0	11.67	0.03
3538	Financial accounts managers	0	9.05	0.22
3539	Business and related associate professionals n.e.c.	0	31.93	0.99
3541	Buyers and procurement officers	0	17.66	0.28
3542	Business sales executives	0	20.56	1.6
3543	Marketing associate professionals	0	2.76	0.13
3544	Estate agents and auctioneers	0	4.27	0.04
3545	Sales accounts and business development managers	0	17.84	1.05
3546	Conference and exhibition managers and organisers	0	3.8	0.03
3550	Conservation and environmental associate professionals	0	20.29	0.01
3561	Public services associate professionals	0	12.54	0.05
3562	Human resources and industrial relations officers	0	4.69	0.24
3563	Vocational and industrial trainers and instructors	0	16.2	0.26
3564	Careers advisers and vocational guidance specialists	0	18.78	0.04
3565	Inspectors of standards and regulations	0	60.79	0.15
3567	Health and safety officers	1	60.95	0.51
4112	National government administrative occupations	0	4.36	0.01
4113	Local government administrative occupations	0	2.87	0
4114	Officers of non-governmental organisations	0	1.27	0
4121	Credit controllers	0	2.33	0.02
4122	Book-keepers, payroll managers and wages clerks	0	0.73	0.02
4123	Bank and post office clerks	0	7.69	0.04
4124	Finance officers	0	0.1	0
4129	Financial administrative occupations n.e.c.	0	2.86	0.03
4131	Records clerks and assistants	0	18.07	0.2
4132	Pensions and insurance clerks and assistants	0	2.89	0.02
4133	Stock control clerks and assistants	0	29.61	0.23
4134	Transport and distribution clerks and assistants	0	15.72	0.14
4135	Library clerks and assistants	0	3.23	0
4138	Human resources administrative occupations	0	0.33	0
4151	Sales administrators	0	1.63	0.02
4159	Other administrative occupations n.e.c.	0	3.88	0.28
4161	Office managers	0	6.66	0.09
4162	Office supervisors	0	11.34	0.15
4211	Medical secretaries	0	1.39	0
4212	Legal secretaries	0	0.05	0
4213	School secretaries	0	1.79	0
4214	Company secretaries	0	1.75	0.02
4215	Personal assistants and other secretaries	0	8.32	0.21
4216	Receptionists	0	0.72	0.01
4217	Typists and related keyboard occupations	0	10.93	0.03
5111	Farmers	0	40.64	0.05

5112	Horticultural trades	0	19.6	0
5113	Gardeners and landscape gardeners	0	10.69	0.03
5114	Groundsmen and greenkeepers	0	13.22	0
5119	Agricultural and fishing trades n.e.c.	0	12.87	0
5211	Smiths and forge workers	1	23.12	0
5212	Moulders, core makers and die casters	1	91.81	0.03
5213	Sheet metal workers	1	79.34	0.07
5214	Metal plate workers, and riveters	1	72.4	0.01
5215	Welding trades	1	98.87	0.5
5216	Pipe fitters	1	98.31	0.01
5221	Metal machining setters and setter-operators	1	96.16	0.7
5222	Tool makers, tool fitters and markers-out	1	91.97	0.14
5223	Metal working production and maintenance fitters	1	90.27	1.17
5224	Precision instrument makers and repairers	1	70.7	0.09
5225	Air-conditioning and refrigeration engineers	0	99.8	0.15
5231	Vehicle technicians, mechanics and electricians	1	78.85	1.61
5232	Vehicle body builders and repairers	1	70.06	0.16
5234	Vehicle paint technicians	0	18.37	0.01
5235	Aircraft maintenance and related trades	0	93.74	0.02
5236	Boat and ship builders and repairers	0	78.05	0.01
5237	Rail and rolling stock builders and repairers	0	61.65	0.01
5241	Electricians and electrical fitters	1	98.79	1.28
5242	Telecommunications engineers	1	95.4	0.99
5244	TV, video and audio engineers	1	82.26	0.02
5245	IT engineers	1	96.31	0.32
5249	Electrical and electronic trades n.e.c.	1	98.47	1.71
5250	Skilled metal, electrical and electronic trades supervisors	0	62.15	0.23
5311	Steel erectors	0	94.43	0.04
5312	Bricklayers and masons	0	30.94	0.05
5313	Roofers, roof tilers and slaters	0	79.03	0.09
5314	Plumbers and heating and ventilating engineers	1	73.49	0.66
5315	Carpenters and joiners	0	56.21	0.51
5316	Glaziers, window fabricators and fitters	0	71.84	0.11
5319	Construction and building trades n.e.c.	0	81.45	0.54
5321	Plasterers	0	33.25	0.03
5322	Floorers and wall tilers	0	46.23	0.04
5323	Painters and decorators	0	4.93	0.02
5330	Construction and building trades supervisors	0	87.19	0.2
5411	Weavers and knitters	0	22.3	0
5412	Upholsterers	0	40.9	0.02
5413	Footwear and leather working trades	0	22.68	0.01
5414	Tailors and dressmakers	0	37.88	0.02
5419	Textiles, garments and related trades n.e.c.	0	43.37	0.02
5421	Pre-press technicians	0	33.54	0.02

5422	Printers	0	24.19	0.04
5423	Print finishing and binding workers	0	32.04	0.02
5431	Butchers	0	9.33	0.01
5432	Bakers and flour confectioners	0	8.46	0.01
5433	Fishmongers and poultry dressers	0	13.91	0
5434	Chefs	0	0.04	0
5435	Cooks	0	1.81	0.02
5436	Catering and bar managers	0	0.94	0
5441	Glass and ceramics makers, decorators and finishers	0	49.16	0.02
5442	Furniture makers and other craft woodworkers	0	41.36	0.03
5443	Florists	0	2.27	0
5449	Other skilled trades n.e.c.	0	52.89	0.33
6121	Nursery nurses and assistants	0	0.05	0
6122	Childminders and related occupations	0	0.08	0
6123	Playworkers	0	4.32	0.01
6125	Teaching assistants	0	0.14	0
6126	Educational support assistants	0	1.85	0.01
6131	Veterinary nurses	0	0.64	0
6132	Pest control officers	0	38.58	0.01
6139	Animal care services occupations n.e.c.	0	5.74	0.01
6141	Nursing auxiliaries and assistants	0	2.1	0.02
6142	Ambulance staff (excluding paramedics)	0	29.74	0.03
6143	Dental nurses	0	0.01	0
6144	Houseparents and residential wardens	0	14.09	0.04
6145	Care workers and home carers	0	1.11	0.05
6146	Senior care workers	0	2.07	0.01
6147	Care escorts	0	2.5	0
6148	Undertakers, mortuary and crematorium assistants	0	3.94	0
6211	Sports and leisure assistants	0	2.94	0.01
6212	Travel agents	0	1.1	0.01
6214	Air travel assistants	0	10.6	0
6215	Rail travel assistants	0	31.68	0.01
6219	Leisure and travel service occupations n.e.c.	0	18.17	0.02
6221	Hairdressers and barbers	0	0.49	0
6222	Beauticians and related occupations	0	3.55	0.01
6231	Housekeepers and related occupations	0	2.43	0.01
6232	Caretakers	0	11.2	0.13
6240	Cleaning and housekeeping managers and supervisors	0	20.03	0.07
7111	Sales and retail assistants	0	9.35	0.38
7112	Retail cashiers and check-out operators	0	9.07	0.01
7113	Telephone salespersons	0	1.66	0.02
7114	Pharmacy and other dispensing assistants	0	0.44	0
7115	Vehicle and parts salespersons and advisers	0	10.6	0.03
7121	Collector salespersons and credit agents	0	33.89	0.08

7122	Debt, rent and other cash collectors	0	22.21	0.08
7123	Roundspersons and van salespersons	0	20.96	0.01
7124	Market and street traders and assistants	0	22.15	0.02
7125	Merchandisers and window dressers	0	4.94	0.02
7129	Sales related occupations n.e.c.	0	14.86	0.71
7130	Sales supervisors	0	16.96	0.52
7211	Call and contact centre occupations	0	4.85	0.12
7213	Telephonists	0	48.85	0.13
7214	Communication operators	0	20.53	0.03
7215	Market research interviewers	0	3.28	0.01
7219	Customer service occupations n.e.c.	0	2.96	0.07
7220	Customer service managers and supervisors	0	10.76	0.08
8111	Food, drink and tobacco process operatives	0	46.16	0.16
8112	Glass and ceramics process operatives	0	30.33	0.01
8113	Textile process operatives	0	42.4	0.35
8114	Chemical and related process operatives	0	49.65	0.32
8115	Rubber process operatives	0	57.63	0.01
8116	Plastics process operatives	0	88.58	0.07
8117	Metal making and treating process operatives	0	42.9	0.13
8118	Electroplaters	0	65.45	0.02
8119	Process operatives n.e.c.	0	65.86	0.02
8121	Paper and wood machine operatives	0	38.83	0.1
8122	Coal mine operatives	0	44.76	0.01
8123	Quarry workers and related operatives	0	56.61	0.05
8124	Energy plant operatives	0	40.94	0.05
8125	Metal working machine operatives	0	88.98	0.66
8126	Water and sewerage plant operatives	0	94.39	0.05
8127	Printing machine assistants	0	28.82	0.03
8129	Plant and machine operatives n.e.c.	0	38.25	0.51
8131	Assemblers (electrical and electronic products)	0	91.38	0.16
8132	Assemblers (vehicles and metal goods)	0	89.81	0.14
8133	Routine inspectors and testers	0	90.51	0.98
8134	Weighers, graders and sorters	0	18.72	0.04
8135	Tyre, exhaust and windscreen fitters	0	76.12	0.08
8137	Sewing machinists	0	64.53	0.07
8139	Assemblers and routine operatives n.e.c.	0	59.85	0.21
8141	Scaffolders, stagers and riggers	0	63.8	0.07
8142	Road construction operatives	0	61.65	0.04
8143	Rail construction and maintenance operatives	1	78.75	0.01
8149	Construction operatives n.e.c.	0	83.65	0.39
8211	Large goods vehicle drivers	0	28.74	0.84
8212	Van drivers	0	23.7	0.38
8213	Bus and coach drivers	0	44.83	0.1
8214	Taxi and cab drivers and chauffeurs	0	16.73	0.01

8215	Driving instructors	0	14.03	0.02
8221	Crane drivers	0	92.75	0.08
8222	Fork-lift truck drivers	0	72.36	0.34
8223	Agricultural machinery drivers	0	31.94	0.01
8229	Mobile machine drivers and operatives n.e.c.	0	88.91	0.34
8231	Train and tram drivers	0	46.71	0.01
8232	Marine and waterways transport operatives	0	31.14	0.02
8233	Air transport operatives	0	43.01	0.01
8234	Rail transport operatives	0	61.18	0.03
8239	Other drivers and transport operatives n.e.c.	0	26.62	0.05
9111	Farm workers	0	11.95	0.01
9112	Forestry workers	0	34.34	0.01
9119	Fishing and other elementary agriculture occupations n.e.c.	0	16.08	0.02
9120	Elementary construction occupations	0	68.9	0.8
9132	Industrial cleaning process occupations	0	45.3	0.1
9134	Packers, bottlers, canners and fillers	0	37.43	0.09
9139	Elementary process plant occupations n.e.c.	0	61.39	0.38
9211	Postal workers, mail sorters, messengers and couriers	0	7.17	0.02
9219	Elementary administration occupations n.e.c.	0	13.27	0.03
9231	Window cleaners	0	6.97	0
9232	Street cleaners	0	14.88	0
9233	Cleaners and domestics	0	8.65	0.09
9234	Launderers, dry cleaners and pressers	0	6.61	0.01
9235	Refuse and salvage occupations	0	67.51	0.13
9236	Vehicle valeters and cleaners	0	34.91	0.01
9239	Elementary cleaning occupations n.e.c.	0	8.31	0
9241	Security guards and related occupations	0	31.86	0.27
9242	Parking and civil enforcement occupations	0	22.46	0.02
9244	School midday and crossing patrol occupations	0	8.56	0.01
9249	Elementary security occupations n.e.c.	0	24.23	0.06
9251	Shelf fillers	0	5.27	0
9259	Elementary sales occupations n.e.c.	0	3.56	0
9260	Elementary storage occupations	0	28.56	0.31
9271	Hospital porters	0	2.07	0
9272	Kitchen and catering assistants	0	1.98	0.04
9273	Waiters and waitresses	0	5.91	0.08
9274	Bar staff	0	0.8	0.01
9275	Leisure and theme park attendants	0	8.72	0.01
9279	Other elementary services occupations n.e.c.	0	6.65	0.04

Notes: STEM density corresponds to the percentage of jobs in an occupation that are STEM. STEM disciplines include Biological/Biomedical, Physical, and Computer Sciences, Technology, Engineering, and Mathematics/Statistics.

Table 16: STEM Disciplines from the CIP classification

CIP code	CIP Standard Major Title
	Biological & Biomedical Sciences
26.0101	Biology/Biological Sciences, General
26.0202	Biochemistry
26.0203	Biophysics
26.0204	Molecular Biology
26.0209	Radiation Biology/Radiobiology
26.0401	Cell/Cellular Biology and Histology
26.0403	Anatomy
26.0406	Cell/Cellular and Molecular Biology
26.0502	Microbiology, General
26.0504	Virology
26.0507	Immunology
26.0702	Entomology
26.08	Genetics
26.0901	Physiology, General
26.0908	Exercise Physiology
26.0911	Oncology and Cancer Biology
26.1001	Pharmacology
26.1004	Toxicology
26.1102	Biostatistics
26.1301	Ecology
26.1303	Evolutionary Biology
26.1305	Environmental Biology
26.1307	Conservation Biology
26.1309	Epidemiology
26.9999	Biological and Biomedical Sciences, Other
	Computer Sciences
11.0103	Information Technology
11.0104	Informatics
11.0202	Computer Programming, Specific Applications
11.03	Data Processing
11.04	Information Science/Studies
11.06	Data Entry/Microcomputer Applications
11.07	Computer Science
11.08	Computer Software and Media Applications
11.0801	Web Page, Digital/Multimedia and Information Resources Design
11.0802	Data Modelling/Warehousing and Database Administration
11.0803	Computer Graphics
11.0899	Computer Software and Media Applications, Other
11.0901	Computer Systems Networking and Telecommunications

11.1001	Network and System Administration/Administrator
11.1003	Computer and Information Systems Security/Information Assurance
11.1005	Information Technology Project Management
11.1099	Computer/Information Technology Services Administration and Management, Other
	Physical Sciences
40.0201	Astronomy
40.0202	Astrophysics
40.0203	Planetary Astronomy and Science
40.0404	Meteorology
40.05	Chemistry
40.0502	Analytical Chemistry
40.0503	Inorganic Chemistry
40.0504	Organic Chemistry
40.0507	Polymer Chemistry
40.0509	Environmental Chemistry
40.06	Geological and Earth Sciences/Geosciences
40.0601	Geology/Earth Science, General
40.0602	Geochemistry
40.0603	Geophysics and Seismology
40.0605	Hydrology and Water Resources Science
40.08	Physics
40.0806	Nuclear Physics
40.0807	Optics/Optical Sciences
40.1001	Materials Science
40.1002	Materials Chemistry
40.9999	Physical Sciences, Other
	Technology
15	Engineering Technology, General
15.03	Electrical Engineering Technologies/Technicians
15.0305	Telecommunications Technology/Technician
15.0399	Electrical and Electronic Engineering Technologies/Technicians, Other
15.04	Electromechanical Instrumentation and Maintenance Technologies/Technicians
15.0401	Biomedical Technology/Technician
15.0499	Electromechanical and Instrumentation and Maintenance Technologies/Technicians, Other
15.0507	Environmental Engineering Technology/Environmental Technology
15.0613	Manufacturing Engineering Technology/Technician
15.0614	Welding Engineering Technology/Technician
15.07	Quality Control and Safety Technologies/Technicians
15.0701	Occupational Safety and Health Technology/Technician
15.0702	Quality Control Technology/Technician
15.08	Mechanical Engineering Related Technologies/Technicians
15.0803	Automotive Engineering Technology/Technician
15.1102	Surveying Technology/Surveying

15.1202	Computer Technology/Computer Systems Technology
15.1204	Computer Software Technology/Technician
15.1302	CAD/CADD Draughting and/or Design Technology/Technician
15.1306	Mechanical Draughting and Mechanical Draughting CAD/CADD
15.1399	Draughting and Design Technology/Technician, General
15.1503	Packaging Science
	Engineering
14	ENGINEERING
14.02	Aerospace, Aeronautical and Astronautical Engineering
14.03	Agricultural Engineering
14.04	Architectural Engineering
14.0501	Bioengineering and Biomedical Engineering
14.0701	Chemical Engineering
14.0801	Civil Engineering, General
14.0803	Structural Engineering
14.0804	Transportation and Highway Engineering
14.09	Computer Engineering
14.0902	Computer Hardware Engineering
14.0903	Computer Software Engineering
14.1001	Electrical and Electronics Engineering
14.1004	Telecommunications Engineering
14.12	Engineering Physics
14.1801	Materials Engineering
14.1901	Mechanical Engineering
14.2001	Metallurgical Engineering
14.2101	Mining and Mineral Engineering
14.2201	Naval Architecture and Marine Engineering
14.2701	Systems Engineering
14.3301	Construction Engineering
14.3501	Industrial Engineering
14.3601	Manufacturing Engineering
14.3701	Operations Research
14.3801	Surveying Engineering
14.3901	Geological/Geophysical Engineering
14.4201	Mechatronics, Robotics, and Automation Engineering
	Mathematics & Statistics
27.01	Mathematics
27.03	Applied Mathematics
27.0303	Computational Mathematics
27.0305	Financial Mathematics
27.05	Statistics

Table 17: Comparison of occupational distributions, UK 2014

Major SOC Code	Major SOC Name	ASHE	BGT Data
1	Managers, directors and senior officials	9.6	9.9
2	Professional occupations	21.5	28.1
3	Associate professional and technical occupations	14.5	22.5
4	Administrative and secretarial occupations	12.1	9.9
5	Skilled trades occupations	8.0	6.5
6	Caring leisure and other service occupations	9.5	6.6
7	Sales and customer service occupations	8.1	6.2
8	Process, plant and machine operatives	6.0	4.2
9	Elementary occupations	10.7	6.1
	Correlation	0.94	

Notes: Produced by BGT. ASHE is the Annual Survey of Hours and Earnings (ASHE) from the Office for National Statistics (ONS).

Table 18: Comparison of geographic distributions, UK 2014

	Professional Occupations		*Elementary occupations*	
	ASHE	BGT Data	ASHE	BGT Data
EAST MIDLANDS	5.9	5.8	8.9	6.9
EAST OF ENGLAND	8.6	9.2	9.4	11.3
LONDON	18.3	28.0	11.0	16.6
NORTH EAST	3.7	2.2	4.3	2.7
NORTH WEST	10.6	8.8	11.4	9.1
SCOTLAND	9.9	6.0	8.7	8.4
SOUTH EAST	15.1	16.2	12.6	17.9
SOUTH WEST	8.3	7.2	9.0	9.2
WALES	4.6	1.9	5.0	2.4
WEST MIDLANDS	7.8	8.3	10.6	8.9
YORKSHIRE AND THE HUMBER	7.2	6.3	9.1	6.4
Correlation	0.94		0.84	

Notes: Produced by BGT. ASHE is the Annual Survey of Hours and Earnings (ASHE) from the ONS. BGT data normalized for the fact that ASHE does not have data on Northern Ireland Employment, while BGT does.