The London School of Economics and Political Science

Essays in Applied Computational Economics

Inna Grinis

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy

London, April 2017

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of around 60,000 words.

Statement of inclusion of previous work

I confirm that Chapter 1 is a revised version of the paper I submitted for the Master of Research (MRes) degree in Economics at the LSE, awarded in July 2015, and that Chapter 4 is an extended and revised version of the paper I submitted for the Bachelor of Arts degree in Economics at the University of Cambridge, awarded in June 2013.

Abstract

This thesis presents four distinct essays that lie at the intersection of economics and computation.

The first essay constructs an abstract framework for defining skills gaps, mismatches and shortages geometrically and thinking about these phenomena in a unified, formal way. It then develops a job matching model with imperfect information, in which skills mismatches influence the job application decisions of the workers, while skills gaps and shortages shape the competition for workers on the resulting bipartite job applications network. The tools proposed in this chapter could in future work be employed as the main ingredients of an agent-based model used to investigate how skills gaps, mismatches and shortages affect equilibrium outcomes.

The second chapter designs and tests machine learning algorithms to classify 33 million UK online vacancy postings into STEM and non-STEM jobs based on the keywords collected from the vacancy descriptions and job titles. The goal is to investigate whether jobs in "non-STEM" occupations (e.g. Graphic Designers, Economists) also require and value STEM knowledge and skills (e.g. "Microsoft C#", "Systems Engineering"), thereby contributing to the debate on whether or not the "STEM pipeline leakage" – the fact that less than half of STEM graduates in the UK work in STEM occupations - should be considered as highly problematic.

Chapter 3 relates to empirical growth. It proposes a programming algorithm, called "iterative Fit and Filter" (iFF), that extracts trend growth as a sequence of medium/long term average growth rates, and applies it on a sample of over 150 countries. The paper then develops an econometric framework that relates the conditional probabilities of up and down-shifts in trend growth next year to the country's current characteristics, e.g. the growth environment, level of development, demographics, institutions, etc.

Finally, Chapter 4 studies credit risk spillovers in financial networks by modelling default as a multi-stage disease with each credit-rating corresponding to a new infection phase. The paper derives analytical and proposes computer simulation-based indicators of systemic importance and vulnerability, then applies them in the context of the Eurozone sovereign debt crisis.

Acknowledgments

I owe everything to my parents Vadim and Gala, my brother Roland, my grandmother Tamara, and my deceased grandparents Inna, Evgeniya, and Artem. Without you, nothing ever would have been possible and this thesis has no sense for me. Thank you for always believing in me and making me believe in myself.

I am very grateful to John Van Reenen, Anders E. B. Nielsen and Chryssi Giannitsarou for continued encouragements, feedback and supervision of the first two, third and an earlier version of the fourth chapters respectively.

I also thank Philippe Aghion, Erik Baurdoux, Kenneth Benoit, Jon Danielsson, Steve Gibbons, Koen Jochmans, Tatiana Komarova, Stephen Machin, Alan Manning, Guy Michaels, Steve Pischke, Fiona Steele, Silvana Tenreyro, Giacomo Zambelli, Jean-Pierre Zigrand, Global Portfolio Solutions (Goldman Sachs) team members, and participants of the LSE Data Science seminar, the 1st IZA Workshop on the Economics of Education, the 2016 Rimini Conference, the 8th Oxford Education Research Symposium, the 15th SAET Conference, the Systemic Risk Centre research meetings and the LSE Labour work-in-progress seminars for interesting and helpful discussions, comments, and suggestions.

The second chapter of this thesis uses Burning Glass Technologies (BGT) data on UK vacancies. I am very grateful to Hal Bonella, Julia Schreiber and Bledi Taska from BGT for giving me access to this data. Thanks also to Alan Manning for introducing me to Giselle Cory, and to Giselle Cory for introducing me to BGT.

This thesis marks the end of my higher education and is a great opportunity for me to also thank all those who have encouraged me and helped me grow intellectually during my undergraduate years at Cambridge: my directors of studies Aytek Erdil and Elisa Faraglia, and supervisors Arun Advani, James Brugler, Tiago Cavalcanti, Zsoka Koczan, Chris Metcalfe, Alexey Onatskiy, Cristiano Ristuccia, Robert Ritz, Donald Robertson, Robert Rowthorn, Riccardo Trezzi, and Melvyn Weeks.

Last but not least, I would like to thank the LSE Economics PhD Programme Manager Mark Wilbor and Department Manager John Curtis, numerous LSE friends and colleagues, and the LSE Orchestra for understanding, kindness, fun, and music.

My PhD studies were kindly funded by the Economic and Social Research Council.

All views expressed, errors and omissions are mine alone.

Contents

List of Tables

List of Figures

List of Algorithms

1 Skills Diversity in Unity

Abstract

At any point in time, skills gaps, mismatches, and shortages arise because of an imperfect correspondence between the singular sets of skills required by different open vacancies and the unique combinations of capabilities embodied in every job seeker - *skills diversity in unity*. This paper first constructs an abstract framework for defining and thinking about these phenomena in a unified, formal and objective way. The main building block is a discrete skills space in which the locations of vacancies and workers are determined by the vectors of skills characterizing them. We define skills gaps and mismatches as two different distance measures between them, and derive a condition for each vacancy that determines whether or not it experiences a skills shortage. We then develop a job matching model with imperfect information, in which skills mismatches influence the job application decisions of the workers, while skills gaps and shortages shape the competition for workers on the resulting bipartite job applications network. The tools proposed in this paper could in future work be employed as the main ingredients of an agent-based model used to investigate how skills gaps, mismatches and shortages affect equilibrium outcomes in the context of *skills diversity in unity* and imperfect information.

1.1 Introduction

> *"At bottom every man knows well enough that he is a unique being, only once on this earth; and by no extraordinary chance will such a marvelously picturesque piece of diversity in unity as he is, ever be put together a second time."*
>
> Friedrich Nietzsche [129]

There has been a lot of debate around the notions of skills gaps, mismatches, and shortages. Some academics completely deny these issues, for instance disparaging skills gaps as a "zombie idea" (Krugman [113]) or "employer whining" (Cappelli [33]).

At the same time, however, numerous surveys conducted by governmental bodies (e.g. the European Commission [40], the UK Commission for Employment and Skills (UKCES) [96]), lobbying organisations (e.g. the Confederation of British Industry (CBI) [35]), and consulting companies (e.g. KPMG [35], ManpowerGroup [119], Hays and Oxford Economics [91]), have been reporting skills gaps and shortages as main obstacles to business operations and a "handbrake on global growth" (Cox [91]) for years. Both public and private sectors spend large amounts of money on investigating and trying to reduce them. For example, in 2013, J.P. Morgan Chase launched a $250 million initiative *"New Skills at Work"*, with their CEO Jamie Dimon quoting the nearly 11 million unemployed Americans and the concurrent 4 million unfilled jobs as evidence of there being a "gulf between the skills job seekers currently have and the skills employers need to fill their open positions"[48]. Such concerns also play an important role in shaping migration policies - e.g. the UKCES reviews were commissioned by the UK Migration Advisory Committee (MAC). Hence, the evidence is not just a "telephone survey [with] executives" (Krugman [113]).

On the workers' side, many labour economists and special institutions (e.g. the European Centre for the Development of Vocational Training (Cedefop) [68]) have extensively studied and documented skills mismatches, the related phenomena of under and over-education (McGuinness [123], Sattinger [145]), and their dire consequences for wages, job satisfaction, and career prospects (e.g. Allen & Van der Velden [9]). In the UK, for instance, according to the ONS [66], 47% of recent graduates were in non-graduate employment in 2013, and the figure was already high even before the recession, e.g., at 43% in 2007. This constitutes a substantial waste of resources and leads to the following puzzle: why aren't job seekers acquiring the skills needed by employers, thereby eliminating skills gaps, mismatches, and shortages?

The proponents of the idea that skills gaps, mismatches, and shortages are just a "myth" add to the perplexity by saying that if these phenomena did exist in reality, we would observe tight labour market conditions (high wages, low unemployment rates) for those workers who have the scarce skills, but we do not. In a recent article, Shierholz [147] shows evidence for the USA that "unemployed workers dramatically outnumber job openings in all sectors" and "in no occupation is there any hint of wages being bid up in a way that would indicate tight labor markets or labor shortages". According to her, it is therefore "aggregate demand", and not structural skills issues, that is behind the weak job recoveries and high non-graduate employment levels of recent graduates.

Although this view is plausible, it seems to ignore the fact that despite recession-related rises, the numbers and concerns around skills gaps, shortages, and mismatches have been high throughout the business cycle, and this persistence over time might actually be the reason why the "zombie idea [...] refuses to die" (Krugman, [113]).

Perhaps all the debate about the existence and importance of skills gaps, shortages, and mismatches is the result of the ambiguity in their definitions, and the little attention paid to the notion of *skills diversity in unity*. The easiest way to understand the latter idea, is to talk to an actual recruiter who might tell you a story about overqualified applicants not being hired because they exhibited a lack of communication skills during interviews. In general, the "lack" is more substantial, but the key intuition is the same. At any point in time, skills gaps, mismatches, and shortages arise because of an imperfect correspondence between the singular sets of skills required by different open vacancies and the unique combinations of capabilities embodied in every job seeker.

Our paper aims to contribute to this debate by proposing tools, which, in future work, could be employed as the main ingredients of an agent-based model used to investigate how skills gaps, mismatches and shortages affect equilibrium outcomes in the context of skills diversity in unity and imperfect information.

In Section 1.3, we develop an abstract framework for defining skills gaps, mismatches and shortages geometrically and thinking about these phenomena in a unified, formal way. The main building block is a discrete skills space in which the locations of vacancies and workers are determined by the vectors of skills characterizing them. We define skills gaps and mismatches as two different distance measures between them. Conceptualising skills shortages - which occur "when there are not enough people available with the skills needed to do the jobs which need to be done" (British Government's Training Agency [4]) - is more complex. The "not enough" notion implies that skills shortages are not pairwise independent like skills gaps and mismatches. Hence, their existence for different vacancies, and the policies aimed at eliminating them cannot be considered in isolation. This highlights the importance and advantage of using a measurable skills space which directly accounts for interdependencies in a given economy, and provides a clear condition for each vacancy that determines whether or not it experiences a skills shortage. We also show how to determine minimum levels of skills mismatches and skills gaps achievable in an economy if the goal is to simultaneously reduce the number of unmatched agents (unemployed workers and unfilled vacancies).

The second part (Section 1.4) develops a two-sided job matching model with imperfect information, in which skills mismatches influence the job application decisions of the workers, while skills gaps and shortages shape the competition for workers on the resulting bipartite job applications network. A preliminary R code file for simulating the spatial structure and the competitive wage adjustment mechanism is available on request.

1.2 Related Literature

This paper can be related to different areas of theoretical and empirical literature in economics and networks, and management (operations research). In this section, we outline the main differences and similarities, and motivate our modelling approach.

The first papers in economics that pay attention to the notion of skills are Roy [142] and Tinbergen [154], both published in 1951. Tinbergen's discussion of skills heterogeneity is close to the one we present below. He recognizes that "many types of employment [...] require certain abilities in varying degrees", so that "in reality, [...] multi-dimensional descriptions of the nature of occupations [...] have to be considered". However, despite these statements, Tinbergen then assumes that "the nature of the labour required is a one-dimensional quantity", summarized by "one number s", and interpreted as "physical effort". Since Tinbergen intends to investigate the distribution of labour incomes, this simplification seems appropriate and useful in order to get analytical results. However, for the analysis of skills mismatches, gaps, and shortages, it is too restrictive because it leads to ignoring a scenario where people have multiple skills and hence different abilities in performing different jobs. For instance, worker A may be more productive in job X than worker B, while B would be more productive than A in another job Y.

Roy [142] was the first economist to clearly understand this, and also to recognise that this implies self-selection: A would choose to work in job X, while B - in job Y. The main difference between our conception of skills heterogeneity and Roy's, is that in Roy's multiple-index model workers have several types of skills but can only use one skill at a time depending on which occupation they choose. By contrast, we assume that both workers and jobs are characterized by multi-dimensional vectors of skills, and production is decreasing in the skills gap between the skills required by the vacancy and those possessed by the worker.

The main contribution of these two early papers is that they generated a substantial literature in which skills diversity plays a key role.

Assignment models that started with Sattinger [144][1] assume infinite numbers of worker and job types. However, as in Tinbergen [154], heterogeneity in these models is typically defined along one dimension only: from low to high ability for workers, and from easy to complex for jobs/tasks. Another difference with our approach is that these models usually require perfect information about all employers' wage offers and

[1]Teulings & Vieira [153] show how assignment models can be estimated, while Shimer [148] proposes a version with coordination frictions.

all workers' abilities. As discussed below, we assume the worker does not observe the job requirements and the wage offers of all employers; he just applies to a vacancy with some probability that is increasing in the ex-ante (before any network-induced competition for him) utility that he would get if employed in this position. Similarly, employers are unaware of the distribution of workers, but they can perfectly observe the skills of those who do apply for their vacancy (by, for instance, inviting the candidates to interviews).

Despite these differences, assignment models are close to our approach to the extent that skills diversity in them is also the main driving force behind the matching of workers and jobs. In particular, the allocation of workers to jobs in assignment models is governed by heterogeneity in either productivities (more able workers have a comparative advantage in more complex jobs, cf. Sattinger [144]), or preferences (workers have different tastes for performing diverse tasks, cf. Tinbergen [155]). In our model, skills mismatches influence job application decisions, while skills gaps and shortages shape the competition for workers on the resulting bipartite job applications network.

The skills space we use to model skills diversity and define skills gaps, mismatches, and shortages is similar in some respects to the "characteristics" space that forms the basis of hedonic models. Although the first instances of such models were developed to study differentiated products (Lancaster [115], Rosen [137]), the same approach has later been applied to labour markets (e.g. Heckman and Scheinkman [92]). The main idea is that products (workers) are "collections of characteristics" (Lancaster [115]) that yield utility (productive efficiency). By assuming n sectors with different production functions, Heckman and Scheinkman [92] also introduce heterogeneity on the labour demand side. However, by contrast with our approach, they do not directly map their jobs onto the same characteristics space as the one used to conceptualize their workers and do not explicitly model how the different measures of divergences between the skills vectors supplied and those demanded affect utility and production.

Perhaps the most important difference between the models discussed so far and the approach we take in this paper lies in the pricing of skills. In both hedonic and assignment models, the prices of different types of skills (characteristics) are determined by equilibrium between their supplies and demands. Although Heckman and Scheinkman [92] show that "whenever population skill endowments are "diverse" enough", skills bundling matters so that "separate productive attributes" command different prices in different sectors, they do not depart from the assumption that there exists a direct mapping from the characteristics of a person to the wage received. After rejecting empirically the hypothesis of uniform factor prices in US sectoral data, they propose factor immobility and non-linear hedonic pricing as alternative explanations, and even suggest that a linear characteristics pricing approach (Lancaster [115]) could still hold within sectors. The reason we wish to depart from the assumption of a direct mapping from characteristics to the wages observed, even accounting for bundling, is because it leads to an important puzzle in the analysis of skills gaps, mismatches, and shortages. If in equilibrium workers are paid according to the overall marginal product and scarcity of their bundles of skills within sectors, why don't rational people recognize the highly valuable types of sector-specific skills and acquire them, thereby eliminating shortage related arbitrage opportunities? Preferences, timing, costs are of course potential reasons. However, another possible explanation is that the hedonic approach does not fully account for the multifaceted role that skills diversity plays in the job matching process and the formation of competitive wages.

In the spatial competitive labour market model proposed below (Section 1.4), skills shortages can induce competition for those workers who possess the scarce bundles, but not necessarily so. The pricing is done through a completely different mechanism in which workers are not necessarily rewarded according to the value of their marginal products and the scarcity of their skills combinations.

To reach these conclusions, we start by recognising that, in reality, a firm could only hire a worker who has applied for its open vacancy. It is important to understand why workers apply to some jobs and not to others, and why different workers choose to apply to different numbers of vacancies. In particular, our goal is to investigate how skills heterogeneity influences such decisions.

Several approaches have been used in the economic literature to model the application process (first stage of the job matching). Undirected search models (Pissarides [131]) assume that workers and firms meet at random, thereby completely ignoring heterogeneity and the fact that workers should apply with a higher probability to jobs that would potentially give them higher utility. Directed search models (e.g. Moen [126], Galenianos & Kircher [79], Shimer [148]) do take this into account. However, they require workers to be able to observe all job offers, and design application strategies that are optimal given all other agents' strategies. Hence, when applying to jobs, workers must not only have perfect information, but also a certain level of strategic sophistication.

In order to avoid such unrealistic assumptions, we propose a novel approach that takes inspiration from the literature on spatial networks where link formation depends on distances (Janssen [103]). Specifically, we assume that for each worker, there exists a latent ranking of all vacancies which depends on skills mismatches and base wages that together determine ex-ante utility as defined in Section 1.4. The worker does not need to perfectly observe this latent ranking, whose sole function is to determine the probability distribution over the vacancies with which the worker applies to each of them. We also extend this basic set-up to allow for several and different numbers of applications per worker.

The simultaneous application decisions of all workers determine a bipartite network in which a link from a worker to an open vacancy corresponds to a job application. The competitive matching of workers and firms as well as the equilibrium wages depend on this network, as it determines the outside opportunities of both firms and workers. Firms that receive several applications for a given vacancy have to choose one among the different candidates, while job seekers who receive multiple offers can only accept one of them at the end of the negotiations.

Kranton & Minehart [112] provide one of the first analysis of competition on bipartite networks. However, the competitive mechanism devised in their paper would be inappropriate in our context because of heterogeneity. Although they show how competitive equilibrium outcomes are influenced by the whole structure of the bipartite buyer-seller network in which the outside opportunities "depend on the entire web of direct and indirect links"; the good exchanged in the process is homogeneous in the sense that a buyer's valuation for the good does not change depending on the seller from whom he acquires it. Similarly, sellers do not care about the buyers' identities, but only the price they receive for their good.

Instead, to model competition on the bipartite job applications network, we use a two-sided matching model. In existing economic theory, the matching literature which started with Gale & Shapley [78] probably provides the most general way of conceptualising heterogeneity in market-like settings. The competitive wage-adjustment mechanism we propose in Section 1.4 is related to the one in Crawford & Knoer [43]. Within the mechanism design literature, the main contribution of their paper is to recognize that in labour markets, agents' preferences need to be modelled as flexible because they can change over the negotiation process in which salaries adjust competitively. Despite being quite general, their model assumes perfect information, and as the authors argue themselves this is disadvantageous since "imperfect information is an essential characteristic of real labor market". Indeed, two-sided matching models usually involve algorithms that require each agent to be able to rank all agents on the opposite side, which is implausible in settings with large numbers of heterogeneous agents. This perfect information assumption might therefore be one of the main reasons why, despite their attractive and intuitive approach in such environments, two-sided matching models have not been more widely used in studying large real labour markets[2].

Furthermore, even though heterogeneity in two-sided matching models has important implications for equilibrium outcomes, it remains unfounded, i.e., these models do not explain why some worker-firm pairs produce more or less, and yield more or less job satisfaction to the worker. They simply take pair-specific productivity and job satisfaction levels as exogenously given.

We try to address both issues by modelling skills diversity explicitly in Section 1.3, incorporating it directly into agents' preferences and job application decisions in Sections 1.4.1 and 1.4.2 respectively, then building a two-sided matching model on this spatial framework in Section 1.4.3.

Finally, our paper can also be related to the large empirical literature devoted to understanding skills mismatches and the related phenomena of under and over-education (cf., for instance, Sattinger [145] for a very detailed overview of the literature on qualitative mismatches, their causes and consequences), while Section 1.3.4 employs some ideas and techniques from the operations research literature (e.g. Eiselt & Sandblom [52]).

The discussion presented in this section shows that despite the "little attention [conventional] economic theory pays to the notion of skill" (UKCES [96]), often treating labour as a homogeneous good, many economists have actually pondered over skills and skills diversity/heterogeneity. The differences with our approach arise because the models reviewed above had been developed for different purposes.

Unfortunately, no clear and objective definitions of skills gaps, mismatches and shortages exist in the academic literature, where skills shortages, for instance, are often understood as a phenomenon that "causes vacancies to remain open longer" (Haskel & Martin [102]) and 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 [14]). However, in practice, hiring difficulties, unfilled vacancies, wage rises, etc. are all

[2]They have been successfully applied in smaller settings, where agents on both sides can provide a complete ranking of all the agents on the opposite side of the market. See Roth and Sotomayor [139] for a textbook exposition of two-sided matching models and a discussion of their applications in the labour market for medical interns and college admissions mechanisms. Such models have also been used in "repugnant markets", e.g. kidney exchanges (Roth et al. [138]).

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 [119] 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.

In the next section, we shall attempt to start filling this theoretical gap by proposing basic, geometric definitions of skills gaps, mismatches and shortages in a unified setting.

1.3 Modelling and Measuring Skills Diversity in Unity

The purpose of this section is to model skills diversity among workers and vacancies, and propose clear definitions of skills mismatches, gaps and shortages. Governments around the world have been concerned about skills gaps and mismatches and also want to minimize the numbers of unemployed workers and unfilled vacancies. In Section 1.3.4, we show that in an economy with a realistic degree of skills heterogeneity and no perfect coordination between the combinations of skills supplied and demanded, there will always be some positive minimum levels of skills gaps and mismatches if the objective is also to leave as few unmatched agents as possible.

1.3.1 Participants and skills space

The economy is composed of two finite and disjoint sets of open vacancies $\mathcal{V} = \{V_1, V_2, ..., V_M\}$, and job seekers (workers) $\mathcal{S} = \{S_1, S_2, ..., S_N\}$, with cardinalities $|\mathcal{V}| = M$ and $|\mathcal{S}| = N$ respectively. We use i and S_i, j and V_j interchangeably when referring to workers and vacancies respectively.

Consider an n-dimensional discrete skills space Ω, where each element is an $n \times 1$ skills vector $\vec{\omega} =< \omega_1, \omega_2, ..., \omega_n >$. Each component of the skills vector $\omega_l \in [0, \bar{\omega}_l]$, for $l = 1, ..., n$, corresponds to some specific type of skills (e.g. presentation skills, computer skills, teamwork, etc.). We shall assume that ω_l is discrete and varies between 0 (no l-type skills) and $\bar{\omega}_l$ (expert in l-type skills). For some types of skills, ω_l will be a binary variable $\{0, 1\}$ indicating whether or not the type of skills is possessed/required, while in other cases the value of ω_l will summarize the level of proficiency in the type of skills considered. Concisely, $\Omega = C^n \bigcap \mathbb{Z}^n$ where $C^n = [0, \bar{\omega}_1] \times [0, \bar{\omega}_2] \times ... \times [0, \bar{\omega}_n]$ for some positive integers $\bar{\omega}_l$, $l = 1, ..., n$.

We map both workers and vacancies onto this skills space, i.e. each worker $i \in \mathcal{S}$ corresponds to an n-dimensional non-negative vector of skills or capabilities $\vec{s_i} =< s_{1i}, s_{2i}, ..., s_{ni} >$ in Ω. Similarly, each vacancy $j \in \mathcal{V}$ corresponds to an n-dimensional vector $\vec{v_j} \in \Omega$ of skills required to perform job j, i.e. $\vec{v_j}$ is the skills vector that the benchmark candidate for vacancy j would possess. For simplicity, we assume that one firm is responsible for one vacancy only and therefore use the terms vacancy, job, firm, and employer interchangeably.[3]

Note that although we use $\vec{s_i}$ and $\vec{v_j}$ when referring to workers and vacancies respectively, both types of vectors belong to the same skills space Ω. Moreover, the characterization of workers and vacancies could be made as precise as needed by simply increasing the dimensionality of the skills space; e.g. instead of just having "computer skills", we could include more skills types to capture proficiency with different types of computer software. In particular, uniqueness could be reached by setting a degree of heterogeneity $n < \infty$ such that no two workers and no two vacancies are the same. Although this is not necessary, we shall assume such a degree of skills diversity in section 1.4 in order to simplify some of the proofs, leaving the more general case as an extension for future research.

1.3.2 Skills mismatches and skills gaps

Consider an arbitrary job seeker i and an arbitrary vacancy j. As long as the vectors characterizing them in the skills space Ω do not coincide, it is possible to compute a distance between them. We define skills mismatches and skills gaps as two different distance measures:

Definition 1. The ***skills mismatch*** between worker $i \in \mathcal{S}$ and vacancy $j \in \mathcal{V}$, sm_{ij}, is the Euclidean distance on Ω between the vectors $\vec{s_i}$ and $\vec{v_j}$:

$$sm_{ij} = \| \vec{v_j} - \vec{s_i} \| = \sqrt{\sum_{l=1}^{n} (v_{lj} - s_{li})^2} \tag{1.1}$$

[3]A possible extension for future research would be to consider more complex scenarios in which one firm simultaneously opens several vacancies, and can hire workers such as to compensate to some extent the skills deficiencies of ones by skills surpluses of others while still minimizing the overall skills gap of the whole team as defined in eq.1.2.

Since sm_{ij} is a measure of the *overall* distance between worker i and vacancy j, it increases both when the worker is overskilled and when he/she is underskilled in some skill type(s).

Definition 2. The ***skills gap*** between worker $i \in \mathcal{S}$ and vacancy $j \in \mathcal{V}$, sg_{ij}, is a measure of skills *deficiency*:

$$sg_{ij} = \sum_{l=1}^{n} \max\{0, v_{lj} - s_{li}\} \tag{1.2}$$

The skills gap only increases when the worker lacks some of the skills that are necessary for the job ($v_{lj} - s_{li} > 0$).

The pairwise skills mismatches and skills gaps between all workers and vacancies in the economy can be summarized by two NxM matrices **SM** and **SG**. For instance:

$$\mathbf{SG} = \begin{bmatrix} sg_{11} & \cdots & sg_{1M} \\ \vdots & \ddots & \vdots \\ sg_{N1} & \dots & sg_{NM} \end{bmatrix}$$

i.e. the ith row of **SG** records the skills gaps of worker i with all open vacancies, whereas column j of **SG** contains the skills gaps that vacancy j has with respect to all job seekers.

Letting $\delta_l = v_{lj} - s_{li}$, it becomes clear that both sm_{ij} and sg_{ij} are specific cases of a more general distance measure defined as:

$$d_{ij} = \sum_{l=1}^{n} f(\delta_l) \tag{1.3}$$

where $f(.)$ is a monotonically increasing function in δ_l.

Skills gaps and skills mismatches therefore correspond to two different ways of perceiving and measuring divergences between the skills combinations embodied in the workers and those required by the open vacancies. Throughout the paper, we assume that workers care only about skills mismatches, since being employed in a job that matches their skills endowments more closely is both more satisfying and requires less extra effort. At the same time, employers are only concerned with skills gaps because any skills deficiency negatively affects their productivity.

Of course, in reality each agent probably has his/her own subjective perception of skills diversity, and the function $f(.)$ in eq.1.3 could be made agent and/or dimension-dependent to reflect this. For instance, when thinking about skills gaps, an employer might allow over-skills in some dimensions to compensate for under-skills in other ones, or penalize under-skilling in different dimensions differently. However, for simplicity and clarity purposes, we shall focus on skills mismatches and skills gaps as defined in equations 1.1 and 1.2.

A simple example that illustrates why it is important to separate skills mismatches from skills gaps is to consider a higher education graduate and two non-graduate vacancies: a barman and a plumber. Being a barman does not necessitate very specific skills, hence the skills gap between the higher education graduate and the barman vacancy is likely to be very small. However, the skills mismatch may be huge since the higher education graduate won't be able to use many of his skills if employed as a barman. By contrast, consider the higher education graduate and the plumber vacancy. This time, both the skills gap and the skills mismatch are likely to be large if the higher education graduate happens to know nothing about plumbing because a plumber is a non-graduate vacancy that requires specific skills.

1.3.3 Measure space and skills shortages

According to the British Government's Training Agency, a skills shortage occurs "when there are not enough people available with the skills needed to do the jobs which need to be done" (British Government's Training Agency [4]). Using the definitions introduced previously, a worker who has all the skills needed to do a particular job is someone who has a zero skills gap with this job. Eq.1.2 implies that such a *qualified* worker does not necessarily have to match a vacancy's requirements perfectly; he/she can be overskilled in some types of skills. Since a worker can therefore be qualified for many different jobs at the same time, the question of establishing whether or not there are "enough" qualified people available in the economy to "do the jobs which need to be done", i.e. to fill all open vacancies, seems rather non-trivial.

Indeed, contrary to skills mismatches and gaps, which are both measures that are specific to a certain worker-vacancy pair - i.e. i's skills gap and mismatch with vacancy j are unrelated to his/her skills gap and mismatch with a different vacancy h - the question of skills shortages cannot be treated in isolation. Hence, before proposing an objective condition that determines whether or not a vacancy experiences a skills shortage, we need to characterize the measure space of the economy in which vacancies and job seekers co-exist. This shall allow us to model their interdependence and conceptualize the "not enough" notion.

Measure space The distribution of the combinations of skills available in the labour market (skills supply) defines a measure P on Ω. For instance, if the pool of job seekers is such that none of them has a specific combination of skills $\vec{\omega}$, i.e. $\vec{s_i} \neq \vec{\omega}$ for all $i \in \mathcal{S}$, this outcome will have measure zero under P, i.e. $P(\vec{\omega}) = 0$. Furthermore, this measure is such that:

$$P(\vec{\omega}) = |\{i \in \mathcal{S} | \vec{s}_i = \vec{\omega}\}| \tag{1.4}$$

where $|.|$ is the cardinality of the subset.

Hence the measure satisfies:

$$P(\Omega) = \sum_{\vec{\omega} \in \Omega} P(\vec{\omega}) = N$$

In a similar way, we can define a measure Q for the vectors of skills demanded to fill open vacancies. If none of the vacancies requires some combination of skills $\vec{\omega}$ - i.e. $\vec{v_j} \neq \vec{\omega}$ for all $j \in \mathcal{V}$, Q shall assign measure zero to this specific skills vector: $Q(\vec{\omega}) = 0$. Again, this measure is such that:

$$Q(\vec{\omega}) = |\{j \in \mathcal{V} | \vec{v}_j = \vec{\omega}\}| \tag{1.5}$$

and satisfies:

$$Q(\Omega) = \sum_{\vec{\omega} \in \Omega} Q(\vec{\omega}) = M$$

Consider the subset of all the job seekers whose skills gaps with vacancy j, as defined in eq.1.2, are zero. Their corresponding skills vectors lie in:

$$Z_j := \{\vec{\omega} \in \Omega | \omega_l \geqslant v_{lj}, \forall l = 1, ..., n\} \tag{1.6}$$

This includes the workers who possess exactly the $\vec{v}_j$ skills vector, as well as those who are overskilled in some type(s) of skills required by vacancy j but underskilled in none of them.

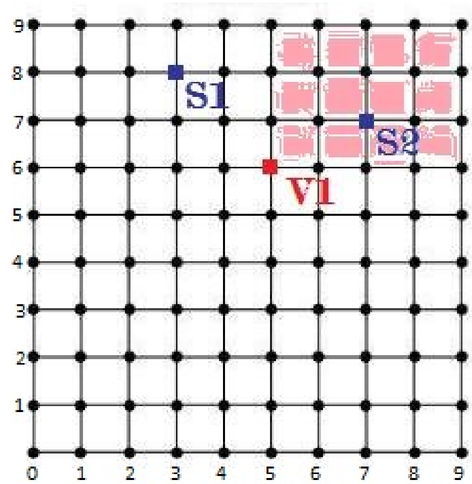
Let us call a job seeker i with $\vec{s}_i \in Z_j$ as *qualified* for vacancy j. Note that, even if the benchmark candidate for vacancy j is absent from the labour force ($P(\vec{v_j}) = 0$), vacancy j might still be able to hire a qualified worker as long as $P(Z_j) > 0$.

Figure 1.1 illustrates the idea in a two-dimensional space. The labour market is composed of one vacancy and two workers characterized by the vectors $\vec{v_1} =< 5, 6 >$,

$\vec{s_1} =< 3, 8 >$, and $\vec{s_2} =< 7, 7 >$ respectively. The shaded area to the North-East of vacancy 1 corresponds to Z_1 as defined in eq.1.6. Only worker 2 ($\vec{s_2}$) belongs to Z_1, despite worker 1 ($\vec{s_1}$) being overskilled for vacancy 1 ($\vec{v_1}$) along the vertical dimension.

Let $\mathcal{Z}$ be the σ-algebra (collection of subsets of Ω) generated by the sets $Z_\omega := \{\vec{u} \in \Omega | u_l \geqslant \omega_l, \forall l = 1, ..., n\}$ for any $\vec{\omega} \in \Omega$.

Figure 1.1 Qualified workers



Notes: The shaded area marks the skills vectors of all the workers who would be qualified for vacancy V1. Only worker S2 has such a combination of skills. S1 is underskilled along the horizontal dimension.

The measure space for this economy is defined as the unique quadruple $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$.

As long as some workers are qualified for several vacancies at the same time, the question of whether or not a given vacancy is experiencing a skills shortage, in the sense of there not being enough qualified job seekers, cannot be addressed by looking at this specific vacancy in isolation. Instead, the vacancy has to be considered within the complete measure space characterizing the economy in which it operates $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$. Both the locations of all the other vacancies and the positions of all the job seekers matter when determining a skills shortage.

Skills shortages

Definition 3. Vacancy j experiences a skills shortage in economy $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$ if:

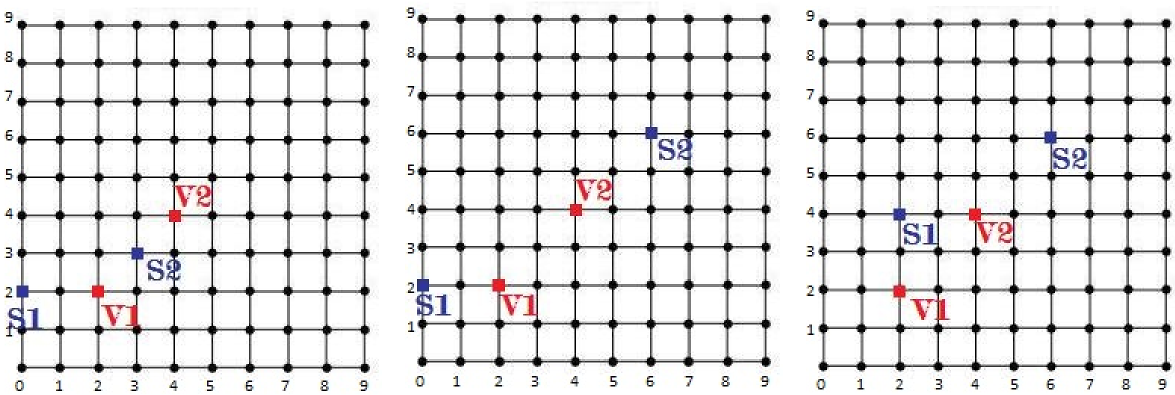
$$Q(\vec{v}_j) + \sum_{\{\vec{\omega} \in \mathcal{H}_j\}} Q(\vec{\omega}) > P(Z_j) + \sum_{\{\vec{\omega} \in \mathcal{L}_j\}} P(\vec{\omega}) \tag{1.7}$$

where $\mathcal{H}_j := \{\vec{\omega} \in \Omega | P(Z_\omega \cap Z_j) \neq 0, \vec{\omega} \neq \vec{v}_j\}$ and $\mathcal{L}_j := \{\vec{\omega} \in \Omega | \vec{\omega} \in Z_u \text{ for } \vec{u} \in \mathcal{H}_j, \text{and } \vec{\omega} \notin Z_j\}$.

The left hand side of eq.1.7 gives the total demand for workers qualified for vacancy j. The first term is j's own demand. The second one sums up the demands from other firms in the same economy that also want to hire workers who are qualified for vacancy j (this is the subset of vacancies with skills vectors in $\mathcal{H}_j$). The first term on the right hand side is the total supply of workers qualified for vacancy j, while the second one adjusts this supply for the fact that firms with skills vectors in $\mathcal{H}_j$, i.e. which compete

Figure 1.2 Skills shortages

Shortage for V2, not for V1 | Shortages for both V1 and V2 | No shortages



Notes: Three possible scenarios with two workers and two vacancies in a two-dimensional skills 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 skills 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. Hence, the existence of shortages and the policies aimed at eliminating them cannot be considered in isolation.

with j for the workers in Z_j, also have access to a pool of workers that are qualified for them but unqualified for vacancy j, and for which they do not compete with j.

To fully understand the condition for a skills shortage contained in eq.1.7, it is useful to look at an example with two dimensions where the problem can be inspected visually.

Figure 1.2 illustrates three possible scenarios with two vacancies and two workers on the same square lattice as the one introduced above.

Let the L, M, and R - subscripts stand for left, middle, and right panels of Figure 1.2. The economy depicted in the left wing panel can be summarized by the quadruple $\Psi^L = \{\Omega, \mathcal{Z}, P^L, Q^L\}$ where Ω is the square lattice, $P^L(\vec{s_1}) = P^L(\vec{s_2}) = 1$ with $\vec{s_1} =< 0, 2 >$, and $\vec{s_2} =< 3, 3 >$ and $P^L(\vec{\omega}) = 0$ for any other $\vec{\omega} \in \Omega$ such that $\vec{\omega} \neq \{\vec{s_1}, \vec{s_2}\}$. For the vacancies, $Q^L(\vec{v_1}) = Q^L(\vec{v_2}) = 1$ with $\vec{v_1} =< 2, 2 >$, and $\vec{v_2} =< 4, 4 >$ and $Q^L(\vec{\omega}) = 0$ for $\vec{\omega} \in \Omega$ such that $\vec{\omega} \neq \{\vec{v_1}, \vec{v_2}\}$. Furthermore, the qualified subsets are such that $P^L(Z_1) = 1$ and $P^L(Z_2) = 0$.

It is clear that vacancy 2 experiences a skills shortage since both potential applicants, S_1 and S_2, lack some of the skills required for V_2's benchmark combination $\vec{v_2}$. Since $P^L(Z_1 \cap Z_2) = 0$, eq.1.7 becomes: $1 + 0 > 0 + 0$ so that vacancy 2 experiences a shortage. By contrast, vacancy 1 does not experience a skills shortage when operating in Ψ^L. The qualified candidate is S_2, and there is *enough* of him/her because he/she is not also qualified for vacancy 2. Eq.1.7 in this case gives $1 + 0 \leq 1 + 0$ since

$Q^L(\vec{v_1}) = P^L(Z_1) = P^L(\vec{s_2}) = 1$ and the second terms on both sides are still equal to 0 because $P^L(Z_1 \cap Z_2) = 0$.

Suppose we change the location of S_2 from $< 3,3 >$ to $< 6,6 >$ while keeping everything else exactly the same as before. This is illustrated in the middle panel of Figure 1.2 . Worker 2 is now the *only* qualified candidate for both vacancies and so there is no longer *enough* of him/her. Indeed, now $P^M(Z_1 \cap Z_2) = 1$ and eq.1.7 becomes $1 + 1 > 1 + 0$ for vacancy 2, and also $1 + 1 > 1 + 0$ for vacancy 1, indicating a skills shortage for both of them.

Finally, the right wing of Figure 1.2 moves job seeker 1 from $\vec{s_1} =< 0,2 >$ to $\vec{s_1} =< 2,4 >$, 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 skills shortages at the level of the economy Ψ^R 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 skills shortage in eq.1.7 is violated for both vacancies. For vacancy 2, the equation reads $1 + 1 \leq 1 + 1$ since $Q^R(\vec{v_2}) = 1$, $P^R(Z_2) = 1$, and $P^R(Z_1 \cap Z_2) = 1$. Note how important it is not to forget the right hand side adjustment $P^R(\vec{s_1}) = 1$. Indeed, although S_2 seems to be over-demanded since he/she is qualified for both vacancies so that total demand for him/her is $Q^R(\vec{v_1}) + Q^R(\vec{v_2}) = 2$, while his/her supply in the economy is only $P^R(\vec{s_2}) = 1$, it would be wrong to conclude that V_2 experiences a skills shortage because S_2 is the only qualified applicant for it. The reason is that, contrary to the situation in Ψ^M, in economy Ψ^R vacancy 1 does have an alternative qualified candidate: S_1. This example illustrates why skills shortages can never be established in isolation because the "not enough" notion is defined relative to the space in which many heterogeneous vacancies and job seekers co-exist.

Another point to note when contrasting the middle and right panels of Figure 1.2 is that simply changing the location of worker 1 eliminates skills shortages for both vacancies. This has an interesting policy implication. In employers' surveys, financial services and engineering firms are often among those that are most concerned with skills shortages, but for different reasons (UKCES [96]). Financial companies often cite the lack of computer/problem solving skills and the lack of understanding of the finance industry as the two crucial deficiencies in their job applicants, while many engineering firms are concerned that their most gifted and best qualified candidates seek finance jobs because of the wage premium this industry is able to pay by leveraging talent (Célérier & Vallée [36]). Is having more graduates majoring in engineering the optimal solution in this case? Maybe not. Top engineers would continue flowing into finance, while lacking some important finance industry knowledge. A more appropriate solution might be to restructure finance/economics degrees so that students majoring in them, who potentially have more of the relevant finance background, also get more computer/problem solving skills. In terms of the spatial framework, this policy would correspond to changing the location of finance/economics graduates in the skills space in order to make them more attractive candidates for the financial industry. This would create more competition to the best engineering graduates, and force some of them back into seeking employment at engineering firms. The example is a caricature of reality, but a useful one to the extent that it illustrates how improving skills shortages for some industry might also alleviate skills shortages in another industry.

A corollary is that when choosing among policies directed at reducing skills shortages in different industries, an authority should include in its costs/benefits analysis the positive externalities that each policy might generate on other industries.

1.3.4 Minimum levels of skills mismatches and skills gaps

We now focus on determining the minimum levels of skills gaps and mismatches achievable in economy Ψ, when the objective is also to match as many workers and vacancies as possible. From a policy perspective, this is an interesting and important question since governments are usually not only concerned with reducing skills gaps and mismatches, but also want to leave as few unemployed workers and unfilled vacancies as possible.

We start with the following definitions:

Definition 4. An ***assignment*** or ***matching*** of workers to firms is a one-to-one correspondence $\mu : \mathcal{V} \bigcup \mathcal{S} \to \mathcal{V} \bigcup \mathcal{S}$ such that:

1. $\mu(j) \in \mathcal{S} \bigcup \{j\}$ for any $j \in \mathcal{V}$;

2. $\mu(i) \in \mathcal{V} \bigcup \{i\}$ for any $i \in \mathcal{S}$;

3. $\mu(j) = i \Leftrightarrow \mu(i) = j$, $j \in \mathcal{V}$ and $i \in \mathcal{S}$, i.e. $\mu(\mu(j)) = j$.

The first two points ensure that a vacancy can either be assigned to a worker in $\mathcal{S}$ or left unfilled (assigned to itself: $\mu(j) = j$). Similarly, a worker can either be assigned to a vacancy in $\mathcal{V}$ or left unemployed (assigned to him/herself $\mu(i) = i$). The last point tells that if a vacancy is assigned to some worker, the worker has to be assigned to this specific vacancy. We will sometimes refer to $\mu(X)$ as the *match* of X for $X \in \mathcal{V} \bigcup \mathcal{S}$.

Moreover, every assignment μ has an associated assignment matrix $\mathbf{A} = [a_{ij}]$ with entries defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } \mu(i) = j \\ 0 & \text{otherwise} \end{cases}$$

Definition 5. Two measures P and Q defined on a sample space Ω are ***equivalent*** if and only if whenever $P(\vec{\omega}) > 0$ we also have $Q(\vec{\omega}) > 0$ and vice versa for any $\vec{\omega} \in \Omega$.

Clearly, given a realistic degree of skills diversity in Ψ, and as long as the formation of skill combinations on the supply side is not perfectly coordinated with the sets of skills demanded, the two probability measures P and Q on Ω will not be equivalent.

If P and Q are not equivalent, at least some of the entries in the skills mismatch matrix $\mathbf{SM} = [sm_{ij}]$, where sm_{ij} is defined in eq.1.1, will be strictly positive. This implies that in an assignment which minimizes the number of unassigned agents, the minimum achievable sum of skills mismatches for the matched pairs - which we denote by SM^{min} - will not necessarily be zero.

To find SM^{min}, we solve a general assignment problem (Kuhn [114]) with an NxM cost matrix $\mathbf{SM}$. Since in general $N \neq M$ the problem is unbalanced. In labour markets, the number of unemployed usually outweighs the number of open vacancies, hence it is plausible to assume that $N > M$. The mathematical problem is then to "pick exactly one element in each [column] (fill each open vacancy) in such a way that each [row] (worker) is used at most once and that the total sum of the [M] elements thus chosen is minimal." (Eiselt & Sandblom [52]). To balance the problem, we introduce $(N-M)$ dummy open vacancies that have zero skills mismatches with all workers. This gives the transformed the NxN skills mismatch matrix $\widetilde{\mathbf{SM}}$. Any worker matched with a dummy vacancy in the final assignment will be considered as unmatched ($\mu(i) = i$).

The general assignment problem with NxN cost matrix $\widetilde{\mathbf{SM}}$ can be solved as the following linear programming problem:

$$\min \sum_{i=1}^{N} \sum_{j=1}^{N} \widetilde{sm}_{ij} a_{ij} \tag{1.8}$$

$$\text{subject to:} \quad \begin{array}{ll} (1) \sum_{i=1}^{N} a_{ij} = 1 & \text{for } j = 1, 2, ..., N \\ (2) \sum_{j=1}^{N} a_{ij} = 1 & \text{for } i = 1, 2, ..., N \\ (3)\ a_{ij} = 1 \text{ or } 0 & \text{for all } i,\ j \end{array}$$

Several algorithms have been proposed to solve this problem. The Hungarian algorithm (Kuhn [114]) is the most famous and earliest one, but many other methods exist (cf. Dell'Amico & Toth [12] for an overview).

There can be several different assignments solving eq.1.8 subject to (1), (2), and (3), but every such assignment μ with assignment matrix $\mathbf{A} = [a_{ij}]$ is ***optimal*** in the sense that:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\widetilde{sm}_{ij}a_{ij} \leq \sum_{i=1}^{N}\sum_{j=1}^{N}\widetilde{sm}_{ij}\hat{a}_{ij} \tag{1.9}$$

for any other assignment $\hat{\mu}$ with assignment matrix $\hat{\mathbf{A}} = [\hat{a}_{ij}]$ which also satisfies constrains (1), (2), and (3).

Since the dummy vacancies introduced have zero skills mismatches with all workers, the minimum skills mismatch achievable in the initial economy characterized by the measure space $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$ such that all M vacancies are filled can be computed as:

$$SM^{min} = \sum_{i=1}^{N}\sum_{j=1}^{N}\widetilde{sm}_{ij}a_{ij} \tag{1.10}$$

If all vacancies are to be filled, a lower level of overall skills mismatches could only be achieved by changing the locations of some agents. However, as soon as at least one worker or vacancy is moved, the measure space and hence the economy change. For a given $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$ and no unmatched vacancies, SM^{min} is therefore the minimum possible sum of skills mismatches for the matched pairs.

From a policy perspective, suppose all vacancies in a given economy Ψ are filled and the actual sum of skills mismatches is SM^R. Knowing SM^{min} will be useful since if $SM^R > SM^{min}$, the policy maker knows that he could achieve a lower overall skills mismatch level by simply reassigning existing workers among existing vacancies, i.e. the assignment itself must be inefficient in terms of skills mismatches. On the other hand, if $SM^R = SM^{min}$, the matching is already optimal since it also satisfies constrains (1), (2), and (3). A lower overall skills mismatch level could only be achieved by implementing policies that change the locations of workers and/or vacancies.

The same exercise could be performed with the skills gap matrix **SG** instead of **SM**. This would yield SG^{min} - the minimum skills gap level achievable in economy $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$ such that all vacancies are filled. From definitions 1 and 2, it is clear that $SG^{min} \leq SM^{min}$ since $sg_{ij} \leq sm_{ij}$ for all i, j. Moreover, if $\mathbf{SG} \neq \mathbf{SM}$, the optimal assignment(s) giving SM^{min} could be very different from those resulting in SG^{min}.

1.4 A Spatial Model of the Competitive Labour Market with Imperfect Information

The previous section created a framework for thinking about skills diversity and the way in which it is perceived by opposite sides of the labour market. We shall now investigate how skills gaps, mismatches, and shortages influence the job matching process and equilibrium outcomes in a competitive labour market with imperfect information. Imperfect information implies that neither workers nor firms are able to perfectly observe the measure space of the economy $\Psi = \{\Omega, \mathcal{Z}, P, Q\}$ in which they operate - an assumption that seems reasonable for a labour market with large numbers of open vacancies and job seekers at any point in time.

We start by determining how skills heterogeneity affects agents' preferences and payoffs, then model the job application and competitive wage adjustment processes, and discuss how skills heterogeneity influences competitive equilibrium outcomes.

1.4.1 Payoffs and rankings

As before, for simplicity, a firm operates exactly one vacancy and can only hire one worker. Similarly, a worker can only be employed in one vacancy.

Let w_{ij} be the wage that worker $i \in \mathcal{S}$ receives if employed in vacancy $j \in \mathcal{V}$. The determination of the competitive wage is discussed below. For this subsection it is enough to think of the wage as some positive real number.

As discussed in section 1.3, we continue to assume that, when considering skills heterogeneity, workers only care about skills mismatches, whereas firms are only affected by skills gaps.

Specifically, the profit of firm j that hires worker $i \in \mathcal{S}$ at wage w_{ij} is given by:

$$\pi_{ij} = p_j y_j(sg_{ij}) - w_{ij} \tag{1.11}$$

where $0 < p_j < \infty$ is the price charged by firm j and $0 \leq y_j(.) < \infty$ is firm j's specific production function which is monotonically decreasing in sg_{ij}. Moreover, let $\pi_{jj} = \psi_j$, i.e. when vacancy j remains unfilled, firm j receives an exogenously given finite amount ψ_j. This amount can either be interpreted as the cost (if $-\infty < \psi_j \leq 0$), or as the present value (if $\infty > \psi_j > 0$) of leaving vacancy j open. In either case, we take ψ_j as given, constant, and known by firm j at any point in time.

Definition 6. At a given wage w_{ij}, worker $i \in \mathcal{S}$ is ***acceptable*** to firm j if and only if $\pi_{ij} = p_j y_j(sg_{ij}) - w_{ij} > \psi_j$. Conversely, a worker is ***unacceptable*** if he/she is not acceptable.

The definition simply says that at a given wage a worker is acceptable to a firm if the firm would prefer to employ the worker at that wage rather than leave the vacancy open. Note that a worker can be acceptable at a given wage w, but become unacceptable at a higher wage $w' > w$.

The utility of worker $i \in \mathcal{S}$ employed in vacancy j at wage w_{ij} is summarized as:

$$u_{ij} = u_i(sm_{ij}, w_{ij}) \tag{1.12}$$

where $\frac{\partial u}{\partial w} > 0$ and $\frac{\partial u}{\partial sm} < 0$. Utility is decreasing in the skills mismatch sm_{ij} because being employed in a job that requires skills further away from his/her own combination of skills is both costlier in terms of effort and less satisfying for the worker. Let $u_{ii} = \kappa_i$ be the utility that a job seeker gets if he/she remains unemployed.

Definition 7. At a given wage w_{ij}, vacancy $j \in \mathcal{V}$ is ***acceptable*** to worker i if and only if $u_{ij} = u_i(sm_{ij}, w_{ij}) > \kappa_i$. Conversely, a vacancy is ***unacceptable*** if it is not acceptable.

Akin to definition 6, definition 7 just tells that a vacancy is acceptable to a worker at a given wage w_{ij} if and only if he/she prefers to be employed in that vacancy at w_{ij} instead of being unemployed and receiving utility $u_{ii} = \kappa_i$. Again, a vacancy can be acceptable to a worker at some wage w, but become unacceptable at a lower wage $w'' < w$.

We also assume that profits and utilities are independent across pairs, i.e. a firm does not directly care about the profits of another firm and a worker's utility is unrelated to the utilities of the other workers.

If worker i observes a given vector of wages $\vec{w_i} =< w_{i1}, w_{i2}, ..., w_{iM} >$ for all open vacancies, and the ith row of the skills mismatch matrix $s\vec{m}_i$, he/she can rank all the open vacancies $\mathcal{V}$ by utility. Let $\mathcal{R}(i, \vec{w_i})$ defined on the set $\mathcal{V} \bigcup \{i\}$ record this ranking. For instance, suppose $M = 3$ and at the given vector of wages $\vec{w_i}$, worker i prefers vacancy V_2 to V_1 and would rather remain unemployed than work for V_3. The worker's ranking can be summarized as: $\mathcal{R}(i, \vec{w_i}) = \{V_2, V_1, S_i, V_3\}$.

Similarly, given a vector of wages $\vec{w_j} =< w_{1j}, w_{2j}, ..., w_{Nj} >$ and the jth column of the skills gap matrix $s\vec{g}_j$, firm j can rank all workers $\mathcal{S}$ in terms of profits. Let $\mathcal{R}(j, \vec{w_j})$ defined on the set $\mathcal{S} \bigcup \{j\}$ record this ranking.

We assume that the rankings $\mathcal{R}(i, \vec{w_i})$ for any $i \in \mathcal{S}$ and $\mathcal{R}(j, \vec{w_j})$ for any $j \in \mathcal{V}$ satisfy the properties of complete ordering and transitivity. For the workers, complete ordering implies that for any two open vacancies V_j and V_k, $j \neq k$, characterized by benchmark skill requirements $\vec{v}_j$ and $\vec{v}_k$ and offering wages w_{ij} and w_{ik} to worker i respectively, the worker can always rank the two vacancies, and say whether or not these vacancies are acceptable to him/her at the given wages. Transitivity implies that if worker i prefers V_j to V_k, and V_k to V_h, $j \neq k \neq h$, at the current wages offered, he/she must also prefer V_j to V_h. The explanations of complete ordering and transitivity are similar on the vacancies' side.

1.4.2 Job application process

The first step in the job matching process is the job application. For a firm to be able to hire a given worker, this worker must not only be seeking employment at the time when the firm opens its vacancy, but he/she must also apply to be considered for this open vacancy before the closing date. It is thus very important to understand what drives application decisions.

The outcomes of the application process can be summarized in a bipartite network. The nodes on the two opposite sides of this bipartite network correspond to the two finite disjoint sets of open vacancies $\mathcal{V} = \{V_1, V_2, ..., V_M\}$ and job seekers $\mathcal{S} = \{S_1, S_2, ..., S_N\}$. A directed link from $\mathcal{S}$ to $\mathcal{V}$ corresponds to a job application. The network therefore records the application decisions of all N workers and is the pictorial representation of the NxM incidence matrix/graph $\mathbf{B}$ with entries defined as:

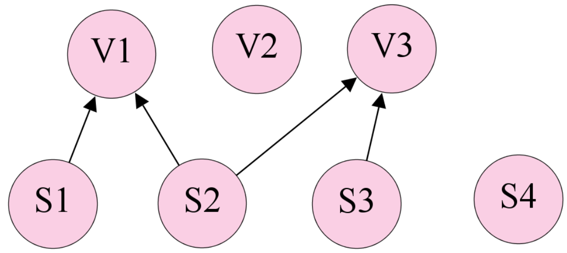
$$b_{ij} = \begin{cases} 1 & \text{if } i \text{ applies to } j \\ 0 & \text{otherwise} \end{cases}$$

As an illustration, Fig. 1.3 shows a simple network with only three vacancies and four workers. The associated incidence matrix is:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Modelling the application process is therefore equivalent to modelling the formation of the links in this bipartite network.

Figure 1.3 Bipartite job applications network



Notes: A link from a worker to a vacancy corresponds to a job application, e.g. S2 applies to both V1 and V3.

Links originate on the workers' side and are directed towards the set of open vacancies. Hence, those links, which lead to potential employment opportunities yielding higher utility levels for the workers, should be relatively more likely to occur. This paper assumes that workers' utility is increasing in wages and decreasing in skills mismatches. A worker should thus be more likely to apply to jobs that pay higher wages and/or with which he/she has a lower skills mismatch.

Specifically, suppose that there exists a base wage w_j for each firm j that can be publicly observed. Either the firm posts this wage together with the benchmark requirements $\vec{v}_j$ when opening its vacancy, or workers can recover it from representative datasets like the Annual Survey of Hours and Earnings (ONS), or websites like *Glassdoor* where people share anonymously the salaries that they are being paid in their jobs.[4] The base wage w_j therefore corresponds to the typical wage paid for this type of job and at this specific firm. Later, we shall also interpret w_j as the wage that a worker receives if employed in vacancy j in equilibrium such that there is no network-based competition for him/her. We take these base wages as given and exogenous, just as a job seeker would do when looking at what wages different firms typically pay.

We refer to the utility evaluated at the base wage as the *ex-ante* utility: $u_{ij} = u_i(sm_{ij}, w_j)$, to differentiate it from the *ex-post* utility which is evaluated at the final competitive wage w_{ij} that i receives in case he/she is matched with j in the competitive equilibrium.

With perfect information about Ψ and w_j, each worker i should be able to compute an *ex-ante* utility $u_{ij} = u_i(sm_{ij}, w_j)$ for all $j \in \mathcal{V}$ and therefore rank all open vacancies. However, in a large real labour market with imperfect information, it would be unreasonable to expect such a complete preference list from any job seeker. The worker

[4] See http://www.glassdoor.co.uk/.

might simply not be able to observe all suitable open vacancies. There could also be some uncertainty about either sm_{ij} or w_j.

To incorporate imperfect information, uncertainty, and unobserved intrinsic preferences into the application process, we propose a model for the formation of the bipartite job applications network that is inspired from the literature on spatial graphs. However, contrary to the standard spatial model in which "nodes are embedded in a metric space" and "link formation depends [only] on the relative position of nodes in the space" (Janssen [103]), we shall assume that link formation is influenced by overall ex-ante utility levels, and therefore not only the skills mismatches but also the base wages.

Intuitively, the probability of a link forming from S_i to V_j in the bipartite network, which corresponds to the probability with which worker i applies to vacancy j, shall be increasing in the utility that i would get if employed at j: $p_{ij} = p_i(u_i(sm_{ij}, w_j))$ with $\frac{\partial p}{\partial u} > 0$. We use the ex-ante utility because the worker cannot foresee the competitive wage he receives in equilibrium. To determine this wage, he/she would need to know how much competition there will be for him/her, if any. However, the latter will depend on the outside opportunities of the firms to which the worker applies - the other applicants for the same vacancies - which the worker cannot observe. Indeed, the worker is unaware of the application decisions of all the other job seekers because he/she does not know the complete measure space Ψ.

To some extent, the spatial labour market model presented here can be seen as a standard spatial model on the skills space, in which the relative positions of the nodes have been deformed by firms being able to pay different base wages. If wages did not enter the utility functions, workers would indeed be most likely to apply to those jobs that match their skills combinations the best. However, anecdotal evidence suggests that the base wage a person expects to be paid in a given job is an important determinant of the application decision. Workers are consciously willing to experience a higher degree of skills mismatch in exchange for a higher wage, and therefore base their application decisions on the overall ex-ante utility they could get in the job considered.

In order to overcome the requirement of perfect information, we assume that given w_j and $\vec{v}_j$ for all $j \in \mathcal{V}$, there does exist for each worker i, a latent ranking of vacancies $\mathcal{R}(i, \vec{w}_i)$. Nevertheless, this ranking does not have to be observed or known completely by the worker, since its sole function is to determine for each worker i, a latent application probability distribution over the set of vacancies $\mathcal{V}$. The only requirement is that application probabilities p_{ij} satisfy for all $i \in \mathcal{S}$ and $j \in \mathcal{V}$:

1. $p_{ij} \in [0, 1]$;

2. if $u_i(sm_{ik}, w_k) > u_i(sm_{ih}, w_h)$, $k \neq h$, then $p_{ik} > p_{ih}$;

3. $p_{ik}(u_i(sm_{ik}, w_k)) = 0$ for any vacancy k that is unacceptable to worker i at w_k.

The resulting bipartite applications network is "self-organizing" because it is "formed by individual actions of autonomous agents" (Janssen [103]). Furthermore, despite being driven by skills mismatches and base wages through their effect on utilities and application probabilities, the job application process remains stochastic, thereby capturing unobserved individual intrinsic preferences for certain positions, uncertainty about either sm_{ij} or w_j, and/or information frictions (with positive probability workers are unaware of the best opportunities available to them in the labour market and hence do not apply for them).

Another characteristic that a realistic application process shall exhibit is a different number of applications per job seeker. Incorporating this feature and understanding why some workers only apply to one or two vacancies, while others apply to many, is important as this plays a crucial role in determining the outside opportunities of the firms and the workers in the competitive adjustment process analysed below. It is reasonable to assume that the cost of applying is decreasing in the skills mismatch (it is easier to write a cover letter for a job that requires your specific combination of skills). Workers are therefore more likely to apply to jobs where their skills mismatch is lower, which just reinforces the effect of skills mismatches on application probabilities that was already present through their effect on utility.

To allow for different numbers of applications per worker, we assume that, for each S_i, the set of links in the bipartite network (the application decisions) is formed by K random draws with replacement from the worker-specific application probabilities distribution over the set of vacancies $\mathcal{V} = \{V_1, V_2, ..., V_M\}$. K is some positive integer and we erase all multiple applications from one worker to the same vacancy. This last step not only yields the desired result of having different numbers of applications per worker, but also the following intuitive insight: those workers for whom there exists a small number of vacancies that provide much higher utility than all the other ones, will on average apply to fewer jobs. In some sense, for these workers there exists one best job profile that acts as a focal point and a larger part of their K applications will be allocated to applying to the job(s) that correspond(s) to this profile. Since any multiple applications are erased, the total number of their applications will be smaller than the average. On the contrary, for those workers who have a relatively general combination of skills and for whom there does not exist one best job profile, application decisions will be more widespread and numerous.

The importance of properly modelling and understanding the application process cannot be overstated. Indeed, competition for workers does not happen at the level of the economy, but exclusively on the bipartite network formed through the simultaneous application decisions of all workers. Hence, the bipartite applications network determines the outside opportunities of both the workers and the firms in the competitive adjustment process that leads to the local (network-based) equilibrium outcomes.

1.4.3 Competitive wage adjustment

In the bipartite applications network, whose formation was studied in the previous subsection, a firm potentially receives zero, one or several applications for its open vacancy. Similarly, a worker applies to zero, one or several open vacancies. The network therefore defines the outside opportunities for all agents. The objective is now to model competition for workers on this bipartite network. We start with several definitions, then propose a competitive wage adjustment process meant to mimic the way in which firms compete for the workers who have applied for their open vacancies.

In this section, in order to simplify some of the proofs, we shall assume that the level of heterogeneity - the n-dimensionality of the skills space Ω - is such that given a wage vector $\vec{w}$, a worker is never indifferent between two separate vacancies, or a vacancy and being unmatched. Similarly, given $\vec{w}$, an employer can always determine whether or not a candidate is acceptable, and for any two acceptable candidates, say which one he prefers.[5]

Definition 8. An ***individually rational outcome*** of the labour market is a one-to-one assignment of workers to vacancies μ and a wage vector $\vec{w}$ such that:

- At $\vec{w}$, workers are acceptable to the vacancies they are assigned to at μ:

$$\pi_{\mu(j)j} = p_j y_j(sg_{\mu(j)j}) - w_{\mu(j)j} \geq \psi_j$$

for all $j \in \mathcal{V}$, and $\pi_{\mu(j)j} = \psi_j$ iff $\mu(j) = j$;

- At $\vec{w}$, vacancies are acceptable to the workers they are assigned to at μ:

[5] A more general treatment could be undertaken in future research, although this assumption is probably not too far away from reality; when comparing two alternatives, it is often possible to find one small extra characteristic that will make us decide in favour of one or the other.

$$u_{i\mu(i)} = u_i(sm_{i\mu(i)}, w_{i\mu(i)}) \geq \kappa_i$$

for all $i \in \mathcal{S}$, and $u_{i\mu(i)} = \kappa_i$ iff $\mu(i) = i$.

Definition 9. An outcome in the ***core*** is an individually rational outcome $(\mu, \vec{w})$ such that no worker-vacancy pair (i, j), with $\mu(i) \neq j$, can negotiate a salary $\widetilde{w}_{ij}$ such that:

- S_i prefers V_j at $\widetilde{w}_{ij}$ to his/her match at $(\mu, \vec{w})$:

$$u_i(sm_{ij}, \widetilde{w}_{ij}) > u_i(sm_{i\mu(i)}, w_{i\mu(i)});$$

- V_j prefers S_i at $\widetilde{w}_{ij}$ to its match at $(\mu, \vec{w})$:

$$\pi_{ij} = p_j y_j(sg_{ij}) - \widetilde{w}_{ij} > \pi_{\mu(j)j} = p_j y_j(sg_{\mu(j)j}) - w_{\mu(j)j}.$$

Definition 10. An assignment μ ***respects*** an incidence matrix $\mathbf{B}$ if for all $i \in \mathcal{S}$ and $j \in \mathcal{V}$: if $\mu(i) = j$ then $b_{ij} = 1$.

In other words, an assignment respects an incidence matrix whenever a firm can only hire a worker who has previously applied for its open vacancy, i.e. there exists a directed link from this worker to the open vacancy in the corresponding bipartite applications network.

Fix an incidence matrix $\mathbf{B}$, formed as described in the previous subsection. The competitive wage adjustment on the corresponding bipartite network is a discrete N-dimensional time process $[\vec{w}_{ij}(t)]_j$, one for each $j \in \mathcal{V}$, with $w_{ij}(t) \in \mathbb{N}$ for any $t = 0, 1, ..., T$, during which wages evolve as follows:[6]

- $t = 0$:

$$w_{ij}(0) = \begin{cases} w_j & \text{if } b_{ij} = 1 \\ \infty & \text{otherwise} \end{cases}$$

i.e. firm j starts by considering all received applications at the base wage w_j. We assume that firm j never makes wage offers below this w_j for some institutional or reputational reasons which we do not investigate here. Given the resulting initial vector of wages $\vec{w}_{ij}(0)$ and the jth column of the skills gap matrix $s\vec{g}_j$, firm j can rank all acceptable applicants, if any, in terms of time zero profits:

[6]Note that the unit of the wage does not matter. It is only necessary that in each time period if a wage rises, the increase is the same constant discrete amount for all firms and workers concerned. Similarly, there is no obvious time interpretation for T. It should simply be conceived as the number of steps necessary for the competitive wage adjustment to converge to an outcome in the core.

$$\pi_{ij}(0) = p_j y_j(sg_{ij}) - w_{ij}(0)$$

Let the firm make a job offer to its best candidate at time $t = 0$, $B_j(0)$, if any, defined as:

$$B_j(0) = \begin{cases} \max_{\{i|b_{ij}=1\}} \pi_{ij}(0) = p_j y_j(sg_{ij}) - w_{ij}(0) & \text{if } \pi_{ij}(0) > \psi_j \\ \emptyset & \text{otherwise} \end{cases}$$

Note that if the firm receives no acceptable applications for its open vacancy, i.e. $\pi_{ij}(0) < \psi_j$ for all i such that $b_{ij} = 1$, it makes no offers at all, $B_j(0) = \emptyset$, leaves the vacancy open and gets $\pi_{jj} = \psi_j > -\infty$.

- $t \in [1, T-1]$:

A worker receives zero, one or several offers. Since in the application process a worker never applies to an unacceptable job and firms cannot offer wages lower than the base wages used by the worker to determine whether or not the vacancy is acceptable to him/her, the worker never receives unacceptable offers. Hence, we condition his/her choice only on the subset of firms which made him/her an offer in the previous step. The worker will *tentatively* hold the best job offer at time t, $H_i(t)$, if any:

$$H_i(t) = \begin{cases} \max_{\{j|B_j(t-1)=i\}} u_{ij}(t) = u_i(sm_{ij}, w_{ij}(t-1)) & \text{if } \{j|B_j(t-1)=i\} \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases} \tag{1.13}$$

The worker *tentatively* rejects the rest of the offers (if any). Until the very last period T all rejections and acceptances of offers are *tentative* because a tentatively rejected firm can come back to the same worker with a higher wage. Indeed, after each round of firm offers and worker decisions, wages adjust as follows:

$$w_{ij}(t) = \begin{cases} w_{ij}(t-1) + 1 & \text{if } B_j(t-1) = i \text{ and } H_i(t) \neq j \\ w_{ij}(t-1) & \text{otherwise} \end{cases} \tag{1.14}$$

i.e. the wage of worker i at firm j rises by one unit at time t only if firm j made an offer to i at $t-1$ and i tentatively rejected this offer.

Given the new vector of wages $\vec{w}_{ij}(t)$, firm j re-optimizes. It recomputes all profits and makes a job offer to its best candidate at time t, $B_j(t)$, if any:

$$B_j(t) = \begin{cases} \max_{\{i|b_{ij}=1\}} \pi_{ij}(t) = p_j y_j(sg_{ij}) - w_{ij}(t) & \text{if } \pi_{ij}(t) > \psi_j \\ \emptyset & \text{otherwise} \end{cases} \tag{1.15}$$

Note that if firm j made an offer at $t-1$ that was tentatively accepted at t, the problem at t is exactly the same as the one faced at $t-1$. Thus, the firm makes the same offer to the same worker at t, i.e. $B_j(t) = B_j(t-1)$. This means that a tentatively accepted offer remains valid until the worker rejects it for another offer that gives him/her higher utility, if he/she ever receives such a better offer before T. If this happens, the problem for the previously tentatively accepted firm changes as the wage of its previous best match increases. The firm re-optimizes and either chooses to come back to the same worker with a higher wage offering if at the new higher wage this worker is still its best alternative, or opts for the new best alternative, which could be leaving the vacancy open in case all workers (including the one that was the best match before the wage rise) become unacceptable at the new wage vector.

- $t = T$:

The competitive wage adjustment process stops when no tentative rejections are issued so that all wages converge: $w_{ij}(T) = w_{ij}(T-1)$ for all $i \in \mathcal{S}$ and $j \in \mathcal{V}$.

The outcome of this competitive wage adjustment process is a matching μ and a wage vector $\vec{w}$, such that:

$$\mu(i) = \begin{cases} j & \text{iff } B_j(T-1) = i \text{ and } H_i(T) = j \\ i & \text{otherwise} \end{cases}$$

$$\mu(j) = \begin{cases} i & \text{iff } B_j(T-1) = i \text{ and } H_i(T) = j \\ j & \text{otherwise} \end{cases}$$

Wages are only defined for the matched pairs, i.e.:

$$w_{ij}(T) = \begin{cases} w_{ij}(T-1) & \text{if } \mu(i) = j \\ \emptyset & \text{otherwise} \end{cases}$$

The payoffs on both sides can be summarized as:

$$u_{ij}(T) = \begin{cases} u_i(sm_{ij}, w_{ij}(T)) & \text{if } \mu(i) = j \\ \kappa_i & \text{otherwise} \end{cases}$$

$$\pi_{ij}(T) = \begin{cases} p_j y_j(sg_{ij}) - w_{ij}(T) & \text{if } \mu(j) = i \\ \psi_j & \text{otherwise} \end{cases}$$

Definition 11. A ***local competitive equilibrium*** on a bipartite network with associated incidence matrix $\mathbf{B}$ is an outcome $(\mu, \vec{w})$, such that $(\mu, \vec{w})$ is in the core and μ respects $\mathbf{B}$.

Theorem 12. *Fix a bipartite network with associated incidence matrix* $\mathbf{B}$. *The competitive wage adjustment process* $[\vec{w}_{ij}(t)]_j$, $j \in \mathcal{V}$, $t = 0, 1, ..., T$, *converges to a local competitive equilibrium in* $T < \infty$ *steps.*

To prove Theorem, we need to show that the outcome resulting from the competitive wage adjustment process $(\mu, \vec{w})$ is individually rational, in the core, and the assignment μ respects the given incidence matrix $\mathbf{B}$. We establish the proof of the theorem through a series of lemmas.

Lemma 13. *A worker who becomes unacceptable to a firm at step* $t < T$, *will never become acceptable to this firm at a later step* $t' > t$.

Proof. If worker i becomes unacceptable to firm j at $t < T$, it must be because $\pi_{ij}(t) = p_j y_j(sg_{ij}) - w_{ij}(t) < \psi_j$. By eq.1.14, at each iteration step in the wage adjustment process, i's wage can either remain constant or rise: $w_{ij}(t') \geq w_{ij}(t)$ for any $t' > t$. This implies that $\pi_{ij}(t') \leq \pi_{ij}(t) < \psi_j$ for any $t' > t$ and completes the proof. □

Lemma 14. *The wage adjustment process results in an assignment* μ *that respects the initially given bipartite applications network with associated incidence matrix* $\mathbf{B}$.

Proof. The wage adjustment process starts by setting $w_{ij}(0) = \infty$ for all workers i such that $b_{ij} = 0$. This implies that any worker who has not applied for vacancy V_j in the initially fixed bipartite network becomes unacceptable already at $t = 0$ since for any such i, $\pi_{ij}(0) = -\infty < \psi_j$. By eq.1.15, a firm never makes an offer to an unacceptable worker and by Lemma 13 a worker that becomes unacceptable to a firm at some $t < T$, never becomes acceptable to this firm at a later $t' > t$. This implies that, during the wage adjustment process, firms only make offers, if any, to workers who had previously applied for their open vacancies, i.e. for which $b_{ij} = 1$. Any resulting assignment of workers to vacancies produced by the wage adjustment process will therefore respect $\mathbf{B}$ by construction. □

Lemma 15. *For any $i \in \mathcal{S}$ and $j \in \mathcal{V}$ such that $b_{ij} = 1$ and worker i is acceptable to firm j at w_j, the competitive wage adjustment process $\{w_{ij}(t)\}$ is bounded above.*

Proof. The process starts by setting $w_{ij}(0) = w_j$ for all $i \in \mathcal{S}$ and $j \in \mathcal{V}$ such that $b_{ij} = 1$. If worker i is acceptable to firm j at w_j, the firm could make one or several different wage offers to i during the wage adjustment process. The maximum wage that firm j could ever offer worker i is $p_j y_j(sg_{ij}) - \psi_j < \infty$. At any higher wage w_{ij}, worker i becomes unacceptable to firm j and by eq.1.15 a firm never makes offers to unacceptable candidates. Furthermore, by eq.1.13, a worker only rejects a previously tentatively accepted offer if he/she receives an offer from a different firm which gives him/her higher utility. Hence, the maximum wage w_{ij}^* that firm j would ever have to offer to worker i such that he/she never rejects its offer for another firm's offer, is such that:

$$u_i(sm_{ij}, w_{ij}^*) = \max_{k \neq j} u_i(sm_{ik}, p_k y_k(sg_{ik}) - \psi_k) + \varepsilon \tag{1.16}$$

where $\varepsilon \to 0$ and k is such that $b_{ik} = 1$ and i is acceptable to k at w_k. Since $p_k y_k(sg_{ik}) - \psi_k$ is finite for all such k, w_{ij}^* is also finite. Hence, for any $i \in \mathcal{S}$ and $j \in \mathcal{V}$, such that $b_{ij} = 1$ and worker i is acceptable to firm j at w_j, the competitive wage adjustment process $\{w_{ij}(t)\}$ is bounded above by

$$\sup w_{ij} = \min\{p_j y_j(sg_{ij}) - \psi_j, w_{ij}^*\} \tag{1.17}$$

where w_{ij}^* is defined by eq.1.16. □

Lemma 16. *The wage adjustment process converges after a final number of steps T.*

Proof. The wage adjustment process converges as soon as no tentative rejections are issued and wages stop rising for all $i \in \mathcal{S}$ and $j \in \mathcal{V}$.

The wage adjustment process is constant for any $i \in \mathcal{S}$ and $j \in \mathcal{V}$ such that $b_{ij} = 0$ ($w_{ij}(t) = \infty$ for all t), or such that $b_{ij} = 1$ and worker i is unacceptable to firm j at w_j ($w_{ij}(t) = w_j$ for all t). This occurs because firm j never makes an offer to such a worker.

Hence, we just need to show that $\{w_{ij}(t)\}$ converges for any $i \in \mathcal{S}$ and $j \in \mathcal{V}$ such that $b_{ij} = 1$ and worker i is acceptable to firm j at w_j. For such (i, j) pairs, eq.1.14 implies that $\{w_{ij}(t)\}$ is monotonically increasing. By Lemma 15, we also know that $\{w_{ij}(t)\}$ is bounded above by $\sup w_{ij}$ as defined in eq.1.17. Since there are finitely many links in the bipartite applications network on which competition for workers can happen, any $\{w_{ij}(t)\}$ will always either converge to its supremum $\sup w_{ij}$ after finitely many steps, or the wage will stop rising at a level below $\sup w_{ij}$ which depends on the amount of competition for worker i in the network. $\square$

Lemma 17. *The outcome $(\mu, \vec{w})$ to which the competitive wage adjustment process converges is in the core.*

Proof. By Lemma 14, μ respects $\mathbf{B}$. It is trivial to show that μ is individually rational. Workers never apply for vacancies that are unacceptable to them at the base wages, and firms never offer wages below the base wages. Hence, μ always assigns workers to either themselves or to vacancies acceptable at $\vec{w}$. On the firms' side, eq.1.15 implies that firms never make offers to unacceptable applicants at any point during the wage adjustment process. This proves that $(\mu, \vec{w})$ is individually rational.

Suppose that $(\mu, \vec{w})$ is individually rational, but not in the core. For this, there must exist a worker-vacancy pair (i, j), with $\mu(i) \neq j$, that can negotiate a salary $\widetilde{w}_{ij}$ such that:

- S_i prefers V_j at $\widetilde{w}_{ij}$ to his/her match at $(\mu, \vec{w})$:

$$u_i(sm_{ij}, \widetilde{w}_{ij}) > u_i(sm_{i\mu(i)}, w_{i\mu(i)}) \tag{1.18}$$

- V_j prefers S_i at $\widetilde{w}_{ij}$ to its match at $(\mu, \vec{w})$:

$$\pi_{ij} = p_j y_j(sg_{ij}) - \widetilde{w}_{ij} > \pi_{\mu(j)j} = p_j y_j(sg_{\mu(j)j}) - w_{\mu(j)j} \tag{1.19}$$

Equation 1.19 implies that at some point in the adjustment process, firm j must have made an offer $\widetilde{w}_{ij}$ to i. By eq.1.15 a firm's decision problem remains the same over time unless the worker rejects its offer. Therefore i must have rejected j's offer at $\widetilde{w}_{ij}$ since otherwise, j would never have made an offer to $\mu(j)$. By eq.1.13, workers reject an offer only if they receive a competing offer that gives higher utility. This implies that i could only have rejected j's offer at $\widetilde{w}_{ij}$ because he/she had a better offer w_{ik} from some firm k at that time. Furthermore, the same argument implies that i's final offer from firm $\mu(i)$ at $w_{i\mu(i)}$ must be at least as good as w_{ik} from k (with equality iff $\mu(i) = k$) :

$$u_i(sm_{i\mu(i)}, w_{i\mu(i)}) \geq u_i(sm_{ik}, w_{ik}) > u_i(sm_{ij}, \widetilde{w}_{ij})$$

which contradicts eq.1.18 and proves Lemma 17.

Lemma 17 completes the proof of Theorem 12. $\square$

1.4.4 Local competitive equilibrium outcomes

The impact of skills heterogeneity on wages for those workers who are matched to vacancies in a local equilibrium can be investigated by looking at the difference between the competitive wage they receive and the base wage for their job: $w_{ij}(T) - w_j$.

First, consider the case when $w_{ij}(T) = w_j$. This could happen for several reasons: worker i applied to one or only a few jobs because his/her utility is very convex in skills mismatches. Any small skills mismatch reduces utility by a lot. For this worker there exists a specific job profile to which he/she applies most of the time, thereby willingly constraining any potential future network-based competition for his/her skills. As a real world example, think about a PhD mathematician who decides to apply only for academic positions, thereby constraining himself any potential competition that there could be for his skills in industry or quantitative finance jobs and that could have pushed his wage above base levels.

Another scenario for $w_{ij}(T) = w_j$ is someone who has a rather general background and applied to many different jobs. His skills gaps are relatively large with most of the vacancies to which he applied so that he is not a candidate for whom there would be a lot of competition (as his wage starts rising he soon becomes unacceptable to many firms to which he applied). This worker receives one offer from some firm that did not get any better candidatures and hires him at the base level.

Finally, $w_{ij}(T) = w_j$ could also happen if i and j are close to a perfect match for each other, so that once i gets an offer from j he never wants to reject it for any other offer, even when j is just offering him the base wage. Potentially j is prepared to compete for i but this never happens because i never rejects j's offer.

On the other hand, $w_{ij}(T) > w_j$ indicates that there was at least one round of network-based competition for the worker, i.e. the worker tentatively rejected the offer of his final employer for a better one at least once in the negotiation process. For this to happen, the worker must himself be of high calibre so that he gets several offers from different employers, but also the worker's final employer must be in a situation where he does not have better alternatives. This could be the case if the vacancy is experiencing a skills shortage as defined in section 1.3.3, so that the rest of the applicants have relatively high skills gaps.

An important insight is that although employers in this model do not care about the profits of other rival firms, they care about both the skills gaps with their own applicants and the skills gaps that the rival firms experience in their applicants' pools. Consider a firm that receives several applications with only one of them being qualified for the vacancy. Even if it is able to hire the qualified candidate in a local equilibrium, larger skills gaps in other candidates imply that the firm will compete for the qualified worker more fiercely since the next best alternatives are not so attractive. If rival firms are in a similar situation and compete for the same qualified worker, his wage could rise well above the base level, eliminating most of the profit for his final employer.

Similarly, it is not necessary for some firm j to be experiencing a skills shortage (as defined in eq.1.7) in order to be induced to participate in network-based competition for its acceptable applicants and potentially lose all of them. The latter could occur if the firms competing with j for its qualified workers are willing and able to raise their wages sufficiently high. Vacancy j could also remain unfilled in a local competitive equilibrium if the workers qualified for j in the economy are relatively unlikely to apply to j. For instance, despite having zero skills gaps, the qualified workers could have substantial skills mismatches with j. Even if they do have relatively low skills mismatches, they could also be poached by other firms that offer higher base wages. Finally, an unfilled vacancy could simply be an unlucky realisation of the random application process.

1.5 Conclusion & Future Research

This paper designs a unified abstract framework that allows us to conceptualise skills gaps, mismatches, and shortages geometrically. We then propose a job matching model meant to mimic the real labour market. In a first step, skills mismatches influence the job application decisions of the workers, who do not have to possess the levels of information and strategic sophistication often assumed in standard economic models reviewed in Section 1.2. Job application decisions result in a bipartite network on which competition, shaped by skills gaps, mismatches and shortages, takes place in a second step.

The skills space, job application and competitive wage adjustment processes can all be simulated as part of an agent-based model, which in future research could be employed to further investigate how skills gaps, mismatches and shortages affect equilibrium outcomes in the context of skills diversity in unity and imperfect information.

Another potential direction for future research is to recreate empirically the measure space for the labour market of higher education graduates, i.e. project real world descriptions of graduates and relevant job openings, together with quantity data, onto a skills space whose dimension will be determined by how detailed the descriptions are. For this, we would need to assemble a detailed list of skills that graduates acquire while studying at university, i.e. create the skills space and construct the skills vectors characterizing higher education graduates. This could be done by looking at specific programme descriptions for each degree and university or using online datasources like LinkedIn, where people often provide detailed information on the courses taken and their own skills. To generate the measure of the combinations of skills supplied (measure P in Section 1.3.3), we would then need to get data on the actual numbers of students graduating in a given year by university and degree. A similar exercise would have to be conducted on the labour demand side, e.g. by using online vacancies data. The measure Q could be constructed from the numbers of job postings that specifically offer graduate employment or do not require substantial amount of work experience.

Once this is done, many interesting questions could be addressed and some policies could be tested employing the techniques developed in Section 1.3. For instance, we could find out objectively (i.e. without asking the employers themselves) which vacancies are experiencing skills shortages, what the minimum achievable levels of skills gaps and mismatches are, etc. This approach would also allow us to experiment with nationwide or university-level curricula reforms, in order to see how they would reduce skills gaps, mismatches, shortages, and improve higher education graduates' employability.

2 The STEM Requirements of "Non-STEM" Jobs: Evidence from UK Online Vacancy Postings and Implications for Skills & Knowledge Shortages

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.

2.1 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 [149]

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.[7] This apparent leakage from the "STEM pipeline" is problematic for two main reasons.

Firstly, it is often considered as the main culprit for the existence of shortages of qualified workers within STEM occupations (e.g. *Scientists, Engineers*), which may have a negative impact on UK's long term growth, e.g. by hindering investments in R&D, technology and innovation (cf. for instance Levy and Hopkins [117], Forth & Mason [69], Harrison [87]).

Secondly, STEM education is more expensive and difficult to acquire than non-STEM one.[8] 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

[7]This finding is robust to different ways of defining STEM occupations and STEM disciplines, e.g. Chevalier [38] examines the Longitudinal Destinations of Leavers from Higher Education (LDLHE) survey 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. [28] find that core STEM occupations employ only 40% of core STEM degree holders.

[8]This is especially true for higher education, which is the main object of interest in this paper since, as we shall see in Section 2.4, most STEM jobs correspond to positions that typically require higher education nowadays. In the UK, part of the total public funding on higher education is provided by the Higher Education Funding Council For England (HEFCE) in the form of grants to universities and colleges. The funds are divided between institutions by using formulae which take into account several factors, including the subjects taught. In particular, there are four distinct price groupings and most STEM subjects fall into the "high cost" and intermediate price categories A to C because they involve some laboratory experiments or fieldwork and are considered as "strategically important" for the UK economy, while the majority of non-STEM subjects belong to price group D which corresponds to "classroom-based subjects". As a concrete example, in 2015-2016, institutions received from the HEFCE between £10,000 per undergraduate studying a subject falling into price category A and £0 per undergraduate enrolled in a subject from category D (cf. HEFCE [54] for more details).

"foundation competencies" (Bosworth et al. [28]), "logical approach to solving problems" (Department for Business, Innovation and Skills (BIS) [125]) 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.[9]

Indeed, "STEM occupations" are a relatively arbitrary construct. They are identified using judgment (Mason [120], BIS [125], BIS [26], Greenwood et al. [84], Department for Innovation, Universities and Skills (DIUS) [65], Chevalier [38]), data-driven approaches (Bosworth et al. [28], Rothwell [141]), or a combination of both (UK Commission for Employment and Skills (UKCES) [64]). Most studies recognize that "the issue of precisely where to draw the line between STEM and non-STEM never goes away" (Bosworth et al. [28]), that "neither Standard Occupational Classification (SOC) system codes nor Standard Industrial Classification (SIC) codes are particularly valuable to [classify STEM employment]" (BIS [125]), and that "STEM degree holders working in a non-STEM occupation may still be using their STEM skills" (Bosworth. et al. [28]).[10]

The only way to shed more light on this important issue is to go beyond *occupations*

[9]For example, see Brynjolfsson & McAfee [31] for a review of how Big Data is transforming management practices.

[10]Mason [120] 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 jobholder's work". The list in Greenwood et al. [84] "was classified by a panel drawn from across the STE subjects and disciplines and convened by The Royal Academy of Engineering". BIS [125] 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. [28] 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 [64] 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 [141] uses O*NET Knowledge scales.

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 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:[11]

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

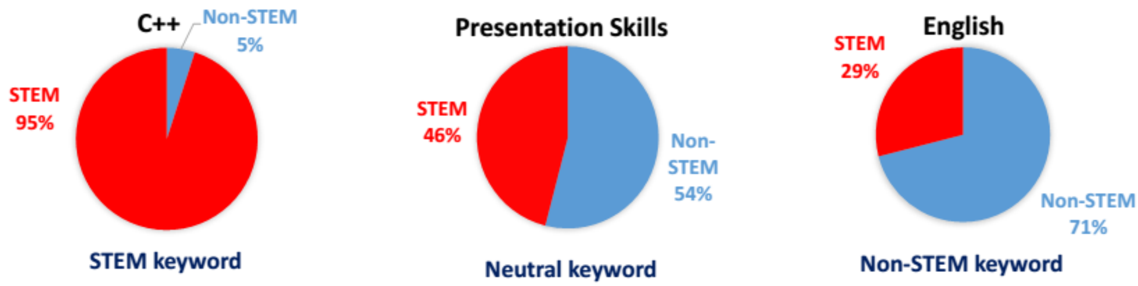
"Financial Analysis – Photography – Rehabilitation"

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. [28]). 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. As we shall see below (Section 2.2), this small subset of postings with explicit discipline requirements is actually very similar to the rest of the vacancies in other respects (e.g. occupational and geographic distributions, keywords posted, etc.). The fact that so few postings contain explicit discipline requirements therefore 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, and not because the 12% of vacancies with explicit discipline requirements form an inherently distinct subgroup.

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 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" (Section 2.3.1). The key idea in "context mapping", illustrated in Figure 2.1, 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.

[11]Figure 2.11 in the Appendix, taken from Carnevale et al. [34], shows how such information is located and extracted from a specific US job advert.

Figure 2.1 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. See Section 2.3.1 for more details.

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.[12] 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 (Section 2.3.2).

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* (i.e. jobs classified as STEM based on the keywords) belong to non-STEM occupations. Of course, nothing prevents STEM graduates working in non-STEM occupations from choosing 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%).

Perhaps surprisingly for the literature where financial occupations are typically considered as the main non-STEM group poaching STEM graduates, none of them is actu-

[12]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".

ally 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 [38], *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 [117]). The evidence often mentioned is the Department for Innovation, Universities and Skills (DIUS) report [65] 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".[13]

In reality, the DIUS report uses the Longitudinal Destinations of Leavers from Higher Education (LDLHE) and the Labour Force surveys and therefore cannot distin-

[13] A similar conclusion is reached by Chevalier [38] 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. [84] 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."

guish 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, we find that STEM jobs are associated with higher wages within both 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 from taking 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 [15] 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* (i.e. 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 [140]), 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 2.5, 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 2.2, 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 [134], 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".[14]

In the US, the paper most related to our work is Rothwell [140]. 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. Contrary to our paper, 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 [141] that we discuss in some details in Section 2.4. 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. [22] 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 [21], the upskilling that happened during the Great Recession has been reversed as the labour market improved from 2010-2014. By contrast, Hershbein and Kahn [93] argue that Ballance et al. [21] "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.

Deming and Kahn [44] also use US BGT data, but this time the goal is to relate variation in skill demands to firm performance and wage variation within occupations.

Although the topic of our paper is quite different from what Hershbein and Kahn

[14] http://wheretheworkis.org/

[93] and Deming and Kahn [44] investigate, it is important to contrast the way in which we work with the BGT data to what these authors have done, in particular because one of the main contributions of our paper is to show how Natural Language Processing (NLP) and Machine Learning (ML) techniques can be employed in exploring the BGT data.

Hershbein and Kahn [93] use the keywords from the BGT vacancies 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 [93] 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 [93] 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 *"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 [44] take a similar approach but go a bit further. 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 [44]).

Neither Hershbein and Kahn [93], nor Deming and Kahn [44] 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.

Our paper lies at the intersection of Data Science and Labour Economics/Economics of Education, and therefore mainly relates to the emerging literature that applies ML and NLP techniques to analyse complex big data and draw novel insights for important policy issues. ML consists of "flexible, automatic approaches [...] used to detect patterns within the data, with a primary focus on making predictions on future data" (Chiu [39]). It is becoming an indispensable toolkit for economists working with big data where standard approaches, like simply classifying a selected number of keywords, are unsatisfactory 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. Varian [157], for instance, calls the tools needed "for manipulating and analyzing big data" the "new tricks" for econometrics, and argues that they "should be more widely known and used by economists".

Einav and Levin [51] provide an interesting discussion and some examples of recent uses of big data in economics. An example of ML application in labour economics is Frey and Osborne [73] who 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).

The rest of this paper is structured as follows. We start by introducing the UK BGT data in Section 2.2. In Section 2.3, we design, test and benchmark ML algorithms to classify thousands of keywords and millions of online job postings into STEM and non-STEM.

We then turn to the characteristics of STEM jobs in the UK in Section 2.4. To the best of our knowledge, we provide the first attempt at going beyond the discrete divide of occupations into STEM vs. non-STEM and documenting the *occupational distribution* of STEM jobs. We look at both - how STEM jobs are distributed across occupations and the percentage of jobs within an occupation that are STEM (STEM density). We can think of the latter as a novel continuous way of describing the importance of STEM requirements at the occupation level. In future work, occupational STEM densities could potentially be employed to help get more precise estimates of STEM demand. Currently, benchmark projections of demand for STEM-educated workers (e.g. Wilson [161], Harrison [87]) are typically developed by combining a historical analysis of patterns of employment by discipline from the Labour Force Survey (LFS) with the forecasts of employment growth by broad sectors and major/sub-major occupations produced in the *Working Futures* by the UKCES (cf. [63] for most recent report). However, as noted in Wilson [161], this approach reflects "both supply and demand influences". We hope that our STEM density estimates contained in Table 17 could help future research get rid of the supply influence, and therefore provide a cleaner estimate of STEM demand.

Section 2.4 also examines the geographic distribution of STEM jobs, the wage premium for STEM, and closes by painting a more detailed profile of STEM jobs within "non-STEM" occupations.

In Section 2.5, we design an abstract framework to help us think about the implications of our empirical findings for the existence of STEM shortages in STEM and non-STEM occupations, as well as education policy initiatives that could help alleviate them.

2.2 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).[15]

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.[16]

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.[17] 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 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

[15] See http://www.careerbuilder.com/, https://jobsearch.direct.gov.uk/, https://www.reed.co.uk/, etc.

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

[17] Figure 2.11 in the Appendix, taken from Carnevale et al. [34], shows how such information is located and extracted from a specific US job advert.

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.

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)). Table 15 in the Appendix shows the distributions across major occupational groups for the 2014 UK BGT sample and the 2014 ONS ASHE. Their correlation is 0.94. 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.[18]

In terms of geographic distributions, the correlation at the regional level is also very high: e.g. 0.95 for the 2015 sample (Table 16, Appendix). However, London is over-represented in the BGT sample relative to ONS employment data by 8% pts.

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."

2.2.1 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 phrases from the vacancy

[18]For the US, Carnevale et al. [34] 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.

description texts. Concretely, in the data, the vacancy description text appears as a set of keywords taken out of context, e.g.:[19]

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

These keywords are collected using "a continuously expanding taxonomy" (Carnevale et al. [34]). 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.[20] 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.

2.2.2 Explicit discipline requirements

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

[19]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".

[20]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. [118]).

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 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.

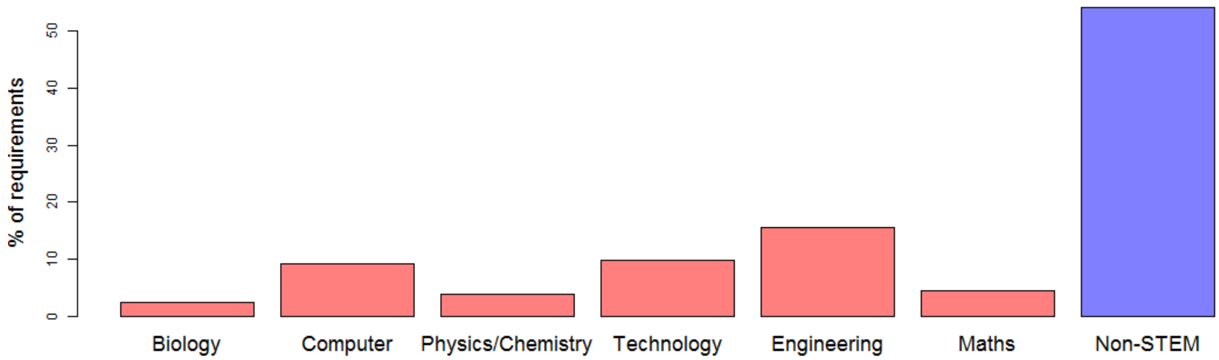
Merging together observations for Jan. 2012 - Jul. 2016, we obtain almost 4 million vacancies with explicit discipline requirements. The 425 distinct disciplines posted in these 4 million vacancies correspond to majors from the *Classification of Instructional Programs* (CIP) - a taxonomic coding scheme 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.[21]

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 18 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.[22]

[21] 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).

[22] 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

Figure 2.2 Distribution of discipline requirements



Notes: 3971988 vacancies with explicit discipline requirements (Jan. 2012 - July 2016).

Around 30% of postings with explicit discipline requirements 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.[23]

Figure 2.2 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.

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.[24] 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

or non-STEM.

[23]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.

[24]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 [64], Mason [120], BIS [26] and Greenwood et al. [84], 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.

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	% of jobs that are STEM	% of STEM jobs
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.

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.

2.3 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 2.3.1, we identify STEM keywords using a method that we call "context mapping". The key idea in "context mapping", illustrated in Figure 2.1, 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 2.3.2 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.

2.3.1 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. 2.2, 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 2.1, 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 2.1 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".[25]

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".[26] 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 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

[25]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).

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

Algorithm 1 Context Mapping

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 [88] 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.

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.[27]

Figure 2.3 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. 2.3 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".[28] *"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 18 in the Appendix).

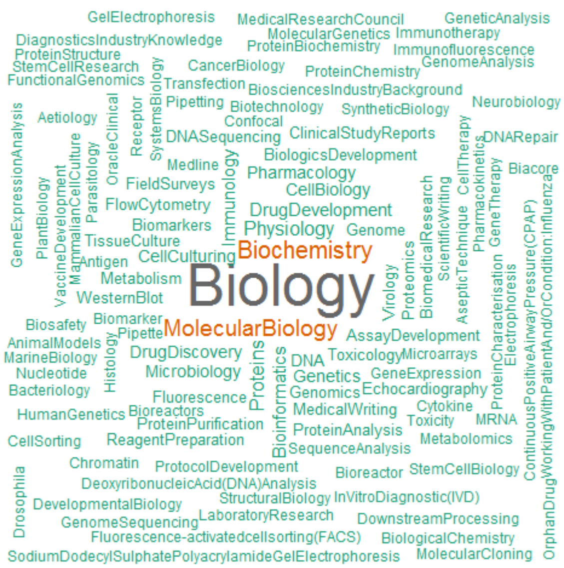
Note that keywords like *"Mathematics"*, *"Computer Skills"*, *"Problem Solving"* all appear in the neutral cluster. This is precisely because within the UK education system, such skills are not exclusively taught in STEM tracks. For instance, *"Mathe-*

[27]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. [98], 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."[98] Similarly, while the objective function in K-means clustering is the residual sum of squares, it would certainly be possible to try different criteria.

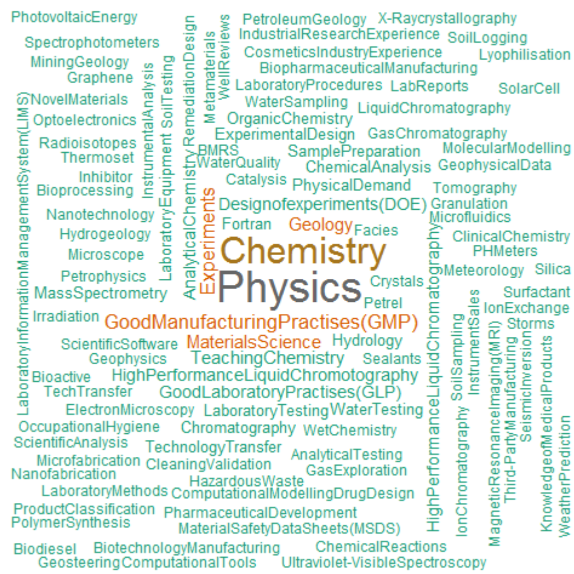
[28]http://w3.siemens.com/mcms/simatic-controller-software/en/pages/default.aspx

Figure 2.3 Examples of STEM (by cluster), non-STEM and neutral keywords

Biological & Biomedical Sciences

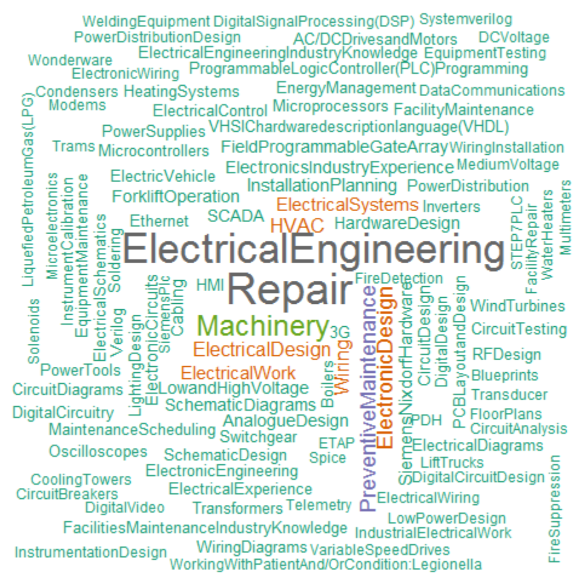


Physical Sciences



Computer Sciences

Technology

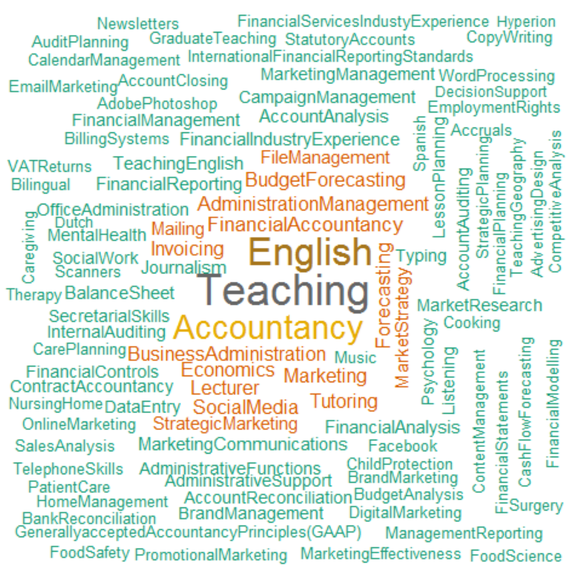


Notes: Continued on next page.

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 [62].

Table 3: Steminess in STEM, Neutral and Non-STEM clusters

Cluster	*Steminess* Mean	Median	Min	*No. Keywords*
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.

matics" 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 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".*

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 &

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.

Biomedical Sciences for the 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.

2.3.2 STEM jobs

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

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.

as an indicator for the fact that its recruiter is going to look preferably for someone with a STEM education.

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

Table 6: Vacancies classification, In-sample performance

Model		*Correctly classified %*	*% non-STEM misclas. into STEM*	*% STEM misclas. into non-STEM*	*Corr. 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.

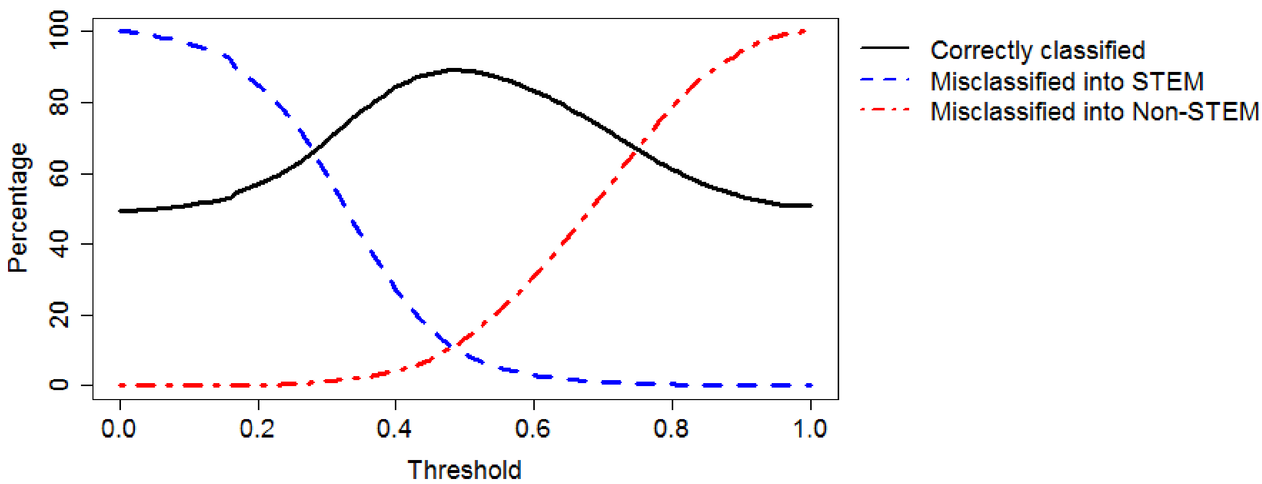
sample with both mixed and non-mixed requirements, the last column of Table 6 shows the correlation of the predicted outcome with the % of STEM discipline requirements. Also note that the tests conducted in Table 6 are *in-sample* because the sample with explicit discipline requirements and keywords ($\mathcal{V}^D \cap \mathcal{V}^K$) 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 2.4 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 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

Figure 2.4 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.

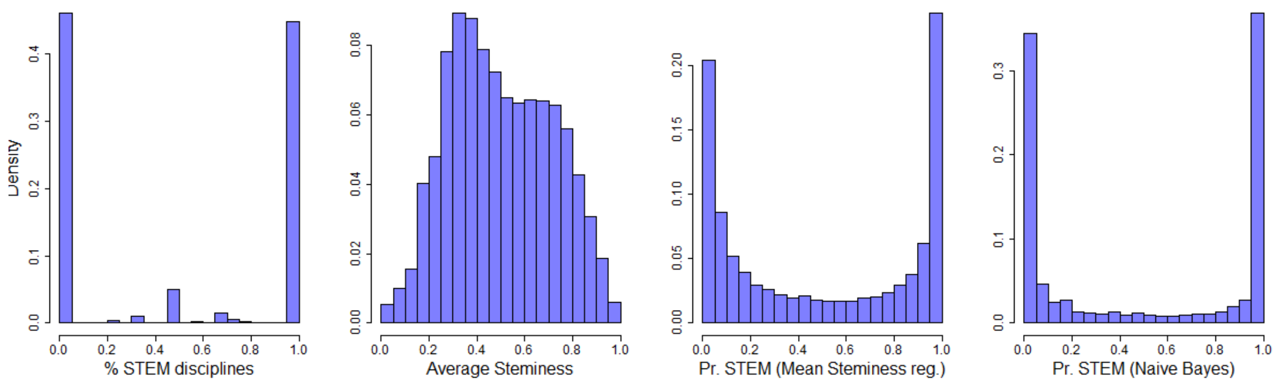
that the technical terms argument may be more important than the precision one. For instance, in model (3) we assign a weight of 1 if a keyword's steminess is computed using at least 50 vacancies. Otherwise the weight is $0.5 + |\mathcal{V}_k|/100$. These weights are then normalized by the total weights' 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. 2.5. 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 logistic link function and estimate the following regression on the sample

Figure 2.5 Comparison of classification strategies with the actual % of STEM disciplines required



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

with non-mixed discipline requirements:[29]

$$\log\left(\frac{\Pr(\text{STEM} \mid steminess_j)}{\Pr(\text{Non-STEM} \mid steminess_j)}\right) = \alpha + \beta steminess_j \tag{2.1}$$

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 $\Pr(\text{STEM} \mid steminess_j) > 0.5$.[30]

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. 2.5 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.2.1 differently. For instance, instead of mean steminess, models (6), (7), and (8) use respectively the per-

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

[30] 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. [101], 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.

centage 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 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:

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

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

$$\Pr(\text{STEM} \mid \mathscr{K}_j^C) = \frac{\Pr(\text{STEM}) \cdot \prod_{k\in\mathscr{K}_j^C} \Pr(k|\text{STEM})}{\prod_{k\in\mathscr{K}_j^C} \Pr(k)} \tag{2.2}$$

$$= \frac{\prod_{k\in\mathscr{K}_j^C} \Pr(\text{STEM}|k)}{\Pr(\text{STEM})^{n_j-1}} \tag{2.3}$$

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

$$\Pr(\text{Non-STEM} \mid \mathscr{K}_j^C) = \frac{\prod_{k\in\mathscr{K}_j^C}(1-\Pr(\text{STEM}|k))}{(1-\Pr(\text{STEM}))^{n_j-1}} \tag{2.4}$$

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. [118], chapters 12 and 13). "Multinomial" because the ordering of the keywords does not matter, "naive" because of the naive assumption of independence.[31] Although this assumption is clearly wrong, in practice, there is simply no way of estimating more complex relationships between keywords given how sparsely

[31] Strictly speaking, the standard implementation of Naive Bayes (NB) uses eq.2.2 and a similar expression for $\Pr(\text{Non-STEM} \mid \mathscr{K}_j^C)$ instead of equations 2.3 and 2.4. 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. [118]), 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 $\Pr(\text{STEM}|k)$.

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

$$steminess_k = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} c_{j,STEM} \tag{2.5}$$

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 $1 - steminess_k$:

$$non - steminess_k = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} c_{j,Non-STEM} \tag{2.6}$$

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. [118], chap. 11). In our case, we can always let the keyword appear in at least one vacancy with perfectly mixed discipline requirements:

$$steminess_k = \frac{1}{|\mathcal{V}_k + 1|} \left\{ \sum_{j \in \mathcal{V}_k} c_{j,STEM} + 0.5 \right\} \tag{2.7}$$

$$non - steminess_k = \frac{1}{|\mathcal{V}_k + 1|} \left\{ \sum_{j \in \mathcal{V}_k} c_{j,Non-STEM} + 0.5 \right\} \tag{2.8}$$

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 2.3 and 2.4 may be above one and not sum to one. However, we can simply normalize them as follows:

$$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) = \frac{\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)}{\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) + \widehat{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C)} \tag{2.9}$$

where:

$$\widehat{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) \quad = \quad \frac{\prod_{k \in \mathscr{K}_j^C} steminess_k}{\widehat{\Pr}(\text{STEM})^{n_j - 1}} \tag{2.10}$$

and

$$\widehat{\Pr}(\text{Non-STEM} \mid \mathscr{K}_j^C) \quad = \quad \frac{\prod_{k \in \mathscr{K}_j^C} (non - steminess_k)}{(1 - \widehat{\Pr}(\text{STEM}))^{n_j - 1}} \tag{2.11}$$

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. 2.5 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.[32]

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

[32]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 (Giga-bytes)*	*% Failed*
(1)	Mean + Max reg.	89.53	9.71	11.26	00:05:35	4.70	0
		[0.134]	[0.198]	[0.191]	[00:00:43]	[0.001]	
(2)	Naive Bayes	89.60	9.22	11.62	00:05:44	4.54	0
		[0.138]	[0.221]	[0.201]	[00:00:48]	[0.001]	
(3)	Logistic Regression with Keywords	87.16	6.39	19.50	04:57:26	14.91	0
		[0.176]	[0.332]	[0.562]	[00:44:20]	[0.046]	
(4)	Linear Discriminant Analysis	89.95	7.77	12.41	08:31:57	95.79	36
		[0.140]	[0.212]	[0.277]	[00:59:47]	[6.645]	
(5)	Support Vector Machines	90.24	6.59	13.04	09:25:42	14.81	2
		[0.128]	[0.211]	[0.237]	[00:51:54]	[0.705]	
(6)	Tree	72.92	2.65	52.26	04:05:38	52.46	8
		[0.410]	[6.578]	[6.725]	[00:36:51]	[0.490]	
(7)	Boosting Tree	77.04	3.03	43.50	05:43:40	56.10	16
		[1.763]	[1.047]	[4.425]	[01:00:04]	[3.308]	
(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. [29]) was used for the implementation of the standard classification methods (models 3-10). As discussed in [29], this package employs a set of optimized algorithms, in particular the *SparseM* package by Koenker and Ng [111]. 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).

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.

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.[33] 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. [101], 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 directly in the logistic regression, the average run of the

[33] Gareth et al. [98] is an excellent introduction into statistical learning. Friedman et al. [101] 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 [61] and the fascinating book on information retrieval by Manning et al. [118].

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. [29] 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 [111] 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. [118] and Friedman et al. [101], 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 the 50% least frequently posted 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. [101]) and is also the reason why many methods (models (8)-(11) in Table 8) simply fail. For instance, we tried k-Nearest Neighbours (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. [101]).

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”* or *“Costing”* 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”* (34.41%) is indeed much more similar to *“Budget Management”* (36.20%) and *“Costing”* (52.28%) than to *“Java”* (90.49%).

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. [98]). 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”.

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. [118]). For instance, tokenizing:

“Uk And Row Process Diagnostic Business Manager”

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.

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 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.).[34]

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

[34]Sanchez [143], Baayen [17] and Feinerer et al. [60] are excellent references on text processing and analysis in R. Natural language processing R packages used in this project include *stringi* (Gagolewski and Tartanus [76]), *stringr* (Wickham [159]), *tm* (Feinerer et al. [59]), *NLP* (Hornik [97]), and *quanteda* (Benoit [24]).

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.

2.4 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 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.

2.4.1 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.

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 24. 2016 includes data up to August only.

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 2.6 which shows the distribution of STEM densities at the four-digit UK SOC level.

Table 17 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 [64], Mason [120], BIS [26] and Greenwood et al.[84] (for the resulting full list of STEM occupations, cf. Footnote 24).

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.[35]

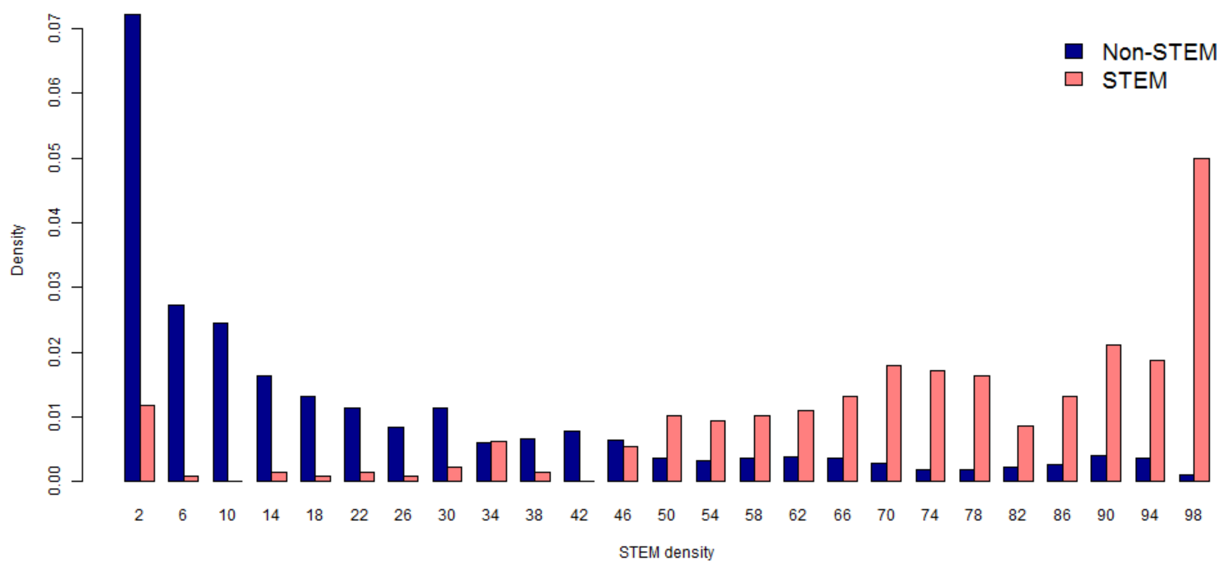
The list of non-STEM occupations with relatively high STEM densities is very diverse. For instance, in 2015, 46.84% of *Business, research and administrative professionals n.e.c.* jobs were identified as STEM. 45.62% of *Product, clothing and related designers,* and even 23.46% of *Artists* looked for STEM graduates. This finding recalls another passage from Matthew Sigelman's inspiring essay on "Why the STEM Gap is Bigger Than You Think" [149] 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 in Chevalier [38], *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

[35] 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 2.6 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.

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 [26] and Mason [120] that investigate vocational STEM skills and apprenticeship training, belong to *Skilled Metal, Electrical and Electronic Trades* (cf. Table 11, fifth 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 [141] 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

Table 11: Occupational distribution of STEM jobs, 2015 (2-digit UK SOC)

Code	*Name*	*STEM density*	*% STEM jobs in*	*% jobs in STEM occ.*	*% All postings*
11	Corporate Managers and Directors	24.31	4.33	15.82	5.82
12	Other Managers and Proprietors	29.59	2.79	0.16	3.07
21	Science, Research, Engineering and Technology Professionals	85.26	39.73	99.77	15.21
22	Health Professionals	1.63	0.24	3.23	4.86
23	Teaching and Educational Professionals	2.99	0.3	0	3.28
24	Business, Media and Public Service Professionals	25.45	6.82	14.8	8.75
31	Science, Engineering and Technology Associate Professionals	76.12	11.46	100	4.91
32	Health and Social Care Associate Professionals	5.92	0.23	15.22	1.26
33	Protective Service Occupations	24.47	0.13	0	0.17
34	Culture, Media and Sports Occupations	15.73	0.98	0	2.03
35	Business and Public Service Associate Professionals	16.04	6.97	1.93	14.18
41	Administrative Occupations	5.93	1.27	0	7.02
42	Secretarial and related Occupations	4.05	0.28	0	2.26
51	Skilled Agricultural and related Trades	20.58	0.09	0	0.14
52	Skilled Metal, Electrical and Electronic Trades	89.79	9.23	94	3.35
53	Skilled Construction and Building Trades	61.92	2.29	24.33	1.21
54	Textiles, Printing and other Skilled Trades	8.45	0.57	0	2.19
61	Caring Personal Service Occupations	1.61	0.18	0	3.67
62	Leisure, Travel and related Personal Service Occupations	7.44	0.26	0	1.15
71	Sales Occupations	12.53	1.88	0	4.91
72	Customer Service Occupations	6.93	0.43	0	2.04
81	Process, Plant and Machine Operatives	60.38	4.74	0.21	2.56
82	Transport and Mobile Machine Drivers and Operatives	35.88	2.25	0	2.05
91	Elementary Trades and related Occupations	57.81	1.4	0	0.79
92	Elementary Administration and Service Occupations	12.21	1.16	0	3.11

Notes: Based on the sample of vacancies with a UK SOC identifier (99.5% of all vacancies posted). Note that the % of jobs in STEM occ. equals zero if none of the 4-digit SOCs in the respective 2-digit subgrouping is typically classified as STEM.

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 [67]). 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".

2.4.2 Spatial distribution

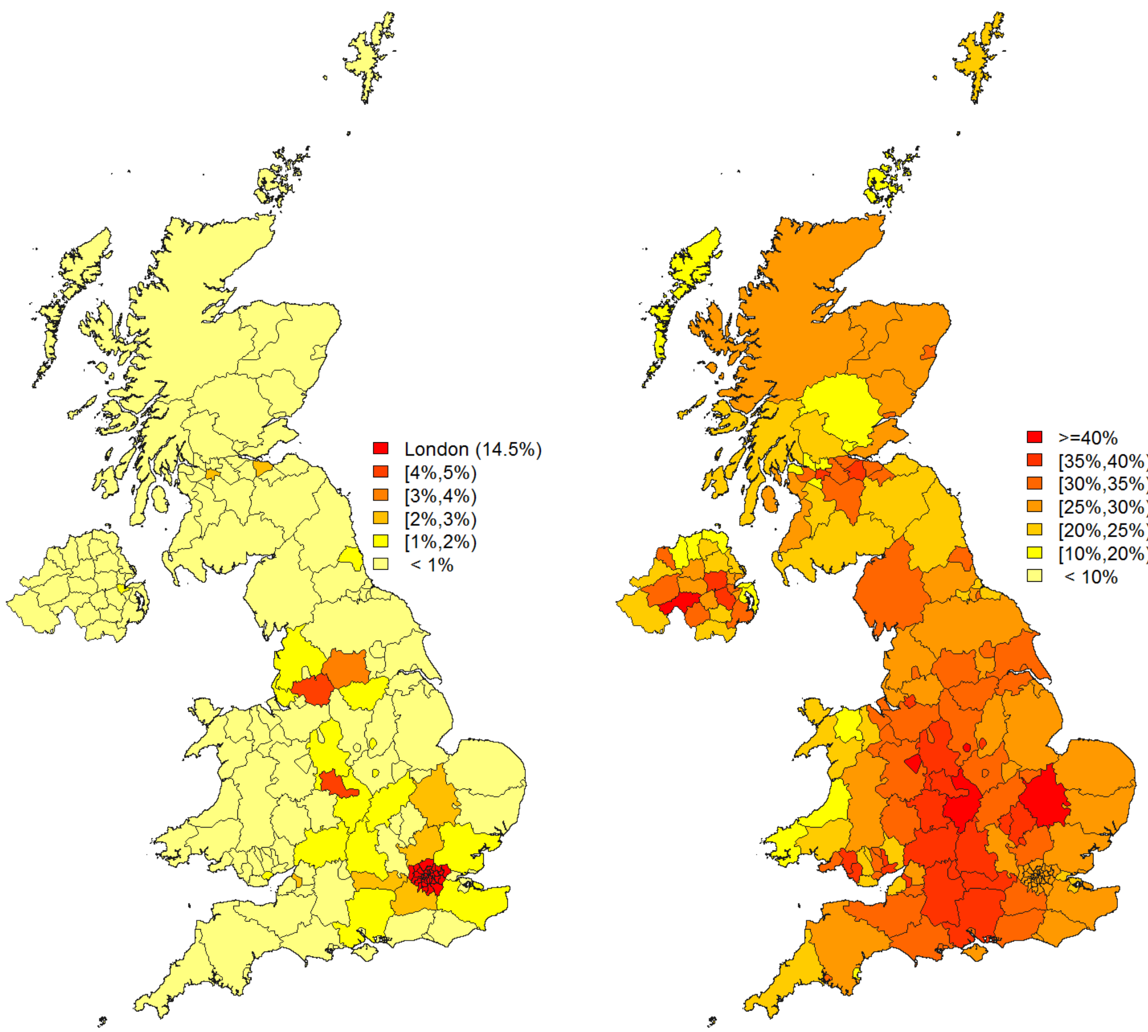
Existing studies indicate that London is a "magnet of STEM workers at the expense of other parts of the country". For instance, Bosworth et al. [28] analyse commuting data and find that London has a net gain of 87,000 Core STEM workers, while the South East, East of England and East Midlands record substantial net losses.

As discussed in Section 2.2, London is over-represented in the BGT sample relative to official employment data (cf. Table 16 in the Appendix), hence we decided to re-weight the spatial distribution of STEM jobs, using the weights shown in the last column of Table 16.[36] The first map in Figure 2.7 shows that even after re-weighting, London 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 14.5% of all STEM vacancies with the next biggest demand for STEM knowledge and skills

[36]The un-weighted map is available on request. Note that weighting only matters for the left map in Figure 2.7. When computing STEM density, weights cancel out (cf. STEM density formula above).

Figure 2.7 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. The left map is re-weighted using the 2015 Annual Survey of Hours and Earnings produced by the Office for National Statistics (cf. Table 16 in the Appendix).

coming from West Midlands (which includes Birmingham and Coventry) with only 4.8% of STEM vacancies, followed by Greater Manchester (4.1%) and West Yorkshire (3.96%). Less than 3% of STEM vacancies were located in any other county.

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. [28], 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.

2.4.3 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:

$$\log w_j = \alpha + \beta STEM_j + \gamma \boldsymbol{X}_j + \varepsilon_j \tag{2.12}$$

$$\log w_j = \alpha + \beta \widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C) + \gamma \boldsymbol{X}_j + \varepsilon_j \tag{2.13}$$

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.[37]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% 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.2.13 instead of eq.2.12. 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,

[37]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.

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.

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. [84], 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."

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)			
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$		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 *STEM occ.							-0.104***		-0.021			
							(0.001)		(0.031)			
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$ *STEM occ.								-0.132***		-0.047		−0.059
								(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.

Other interesting contributions include, for instance, Webber [158] who looks at how average earnings vary by discipline in the US. Bratti et al. [30] use a British cohort study from 1970 to estimate wage returns by major studied. Gabe [75] 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

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)				
$\widetilde{\Pr}(\text{STEM} \mid \mathscr{K}_j^C)$ *STEM occ.			−0.050		−0.037	−0.036
			(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.

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 on all the controls introduced in Tables 13 and 14.[38]

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.

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 [65]). 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.

[38]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.

2.4.4 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 our main object of interest.

In particular, let us 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"...*

Despite the fact that these recruiters are looking for STEM graduates with a higher probability than for non-STEM ones, it seems that 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 occupations from STEM jobs in 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).

This suggests 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).

2.5 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.

2.5.1 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 [102]). 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 [14]).

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 [119] 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) [83] 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 [125], 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".[39]

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 [4], 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.

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.

[39] The academic literature also contains many contributions showing that financial incentives have little or no impact of student learning choices at all education levels, e.g. cf. Fryer [74] 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 $v_{jl} = \phi_j^v \theta_{jl}^v$.

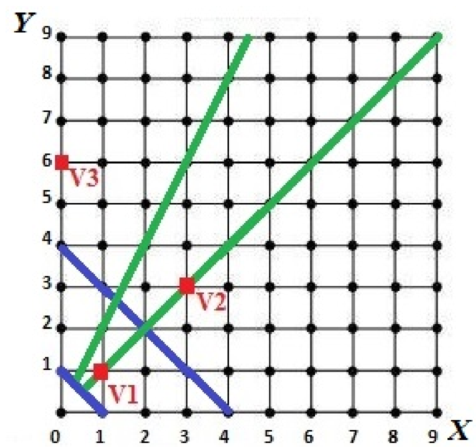
Figure 2.8 illustrates the idea on a two dimensional lattice.

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 2.8 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 18. 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:

$$P(\omega) = |\{i \in \mathcal{S} | s_i = \omega\}| \text{ and } P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = |\mathcal{S}| \tag{2.14}$$

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

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

$$Q(\omega) = |\{j \in \mathcal{V} | v_j = \omega\}| \text{ and } Q(\Omega) = \sum_{\omega \in \Omega} Q(\omega) = |\mathcal{V}| \tag{2.15}$$

Definition 19. Vacancies located at $\Delta \in \Omega$ experience a shortage if:

$$Q(\Delta) + \sum_{\{\omega \in \mathcal{H}_\Delta\}} Q(\omega) > P(Z_\Delta) + \sum_{\{\omega \in \mathcal{L}_\Delta\}} P(\omega) \tag{2.16}$$

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.2.16 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 20. To fully understand the condition for a shortage contained in eq.2.16, we can look at a simple example with a two-dimensional space in which the problem can be inspected visually.

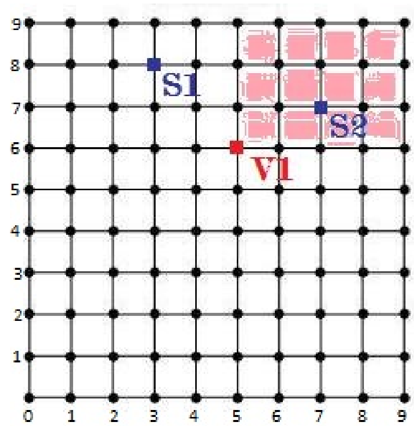
For clarity, let's also assume the following specific form for the qualified subsets, illustrated in Figure 2.9:

$$Z_j = \{\omega \in \Omega | \omega_l \geqslant v_{jl}, \forall l = 1, ..., n\} \tag{2.17}$$

Intuitively, eq.2.17 corresponds to the subset of candidates who have at least as much skills & knowledge in each dimension as what the vacancy requires.

In Figure 2.9, 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.2.17. 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.2.17 because he has more overall skills & knowledge and is located such that $s_{21} > v_{11}$ and $s_{22} > v_{11}$.

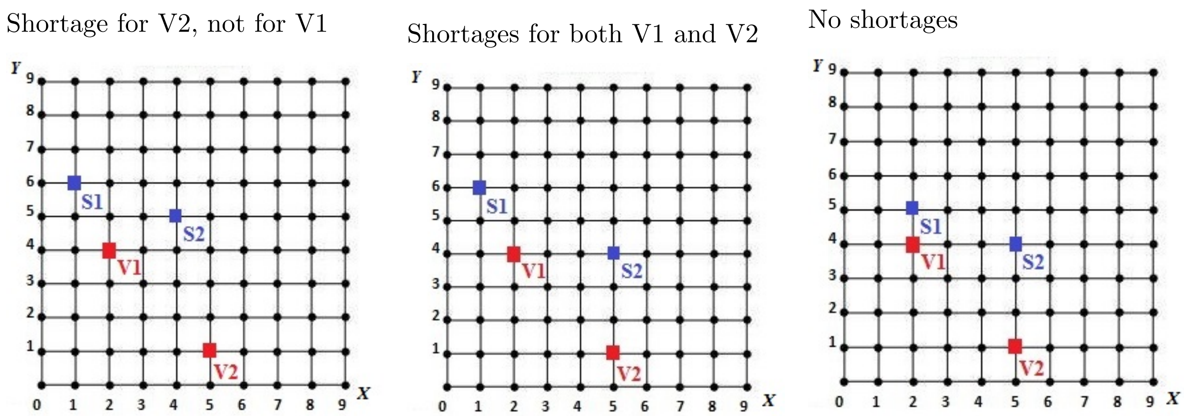
Figure 2.9 Qualified subset



Notes: The shaded area marks the subset of candidates who would be qualified for vacancy V1 according to eq. 1.6

Example 21. Figure 2.10 illustrates how the condition

Figure 2.10 Illustrating the condition for shortages in eq. 1.7



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.

for determining the presence of a shortage in eq. 1.7 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.2.16 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. Eq.2.16 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.2.16 becomes $1 + 1 > 1 + 0$ for both vacancies, indicating shortages.

Finally, in the right wing of fig 2.10 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.2.16 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.

2.5.2 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. 2.10, 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.

2.6 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 so-called "STEM" occupations (e.g. *Scientists, Engineers*) 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 [141] 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.

Other potential extensions of our work include using the occupational STEM density estimates, presented in Section 2.3.2 and Table 17, to get more precise projections of STEM demand, or combining them with the Labour Force Survey to construct an estimate of the percentage of “STEM wastage”, by e.g. subtracting the STEM density estimate (STEM demand) from the percentage of people with STEM degrees working in an occupation (STEM supply).

Another interesting direction is to spend more time investigating the spatial distribution of STEM jobs and relating it to measures of regional development, innovation, unemployment, etc..

Finally, 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”.

2.7 Appendix

Table 15: 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).

Figure 2.11 Extracting information from online job ads (from Carnevale et al. [34])

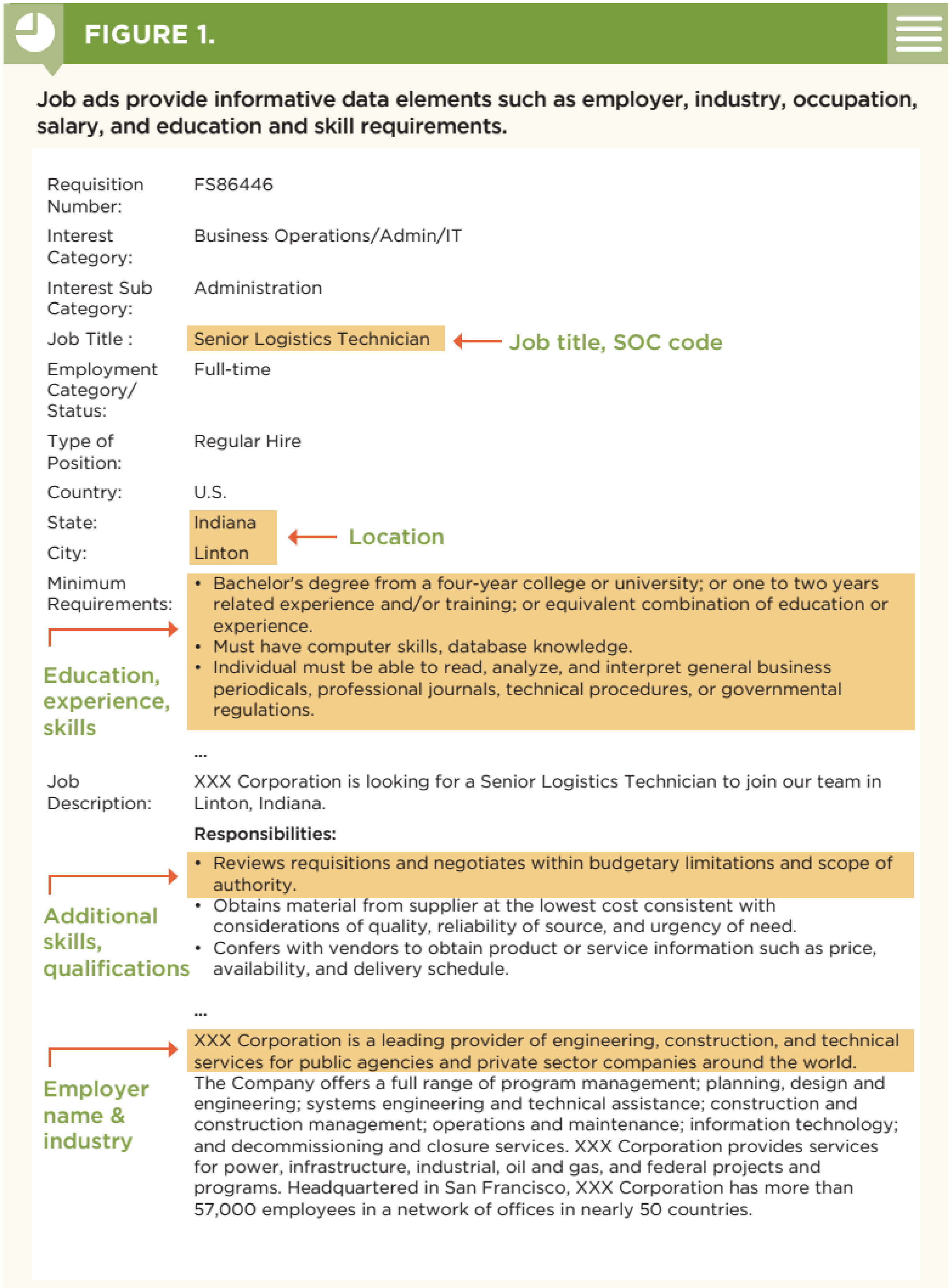


FIGURE 1.

Job ads provide informative data elements such as employer, industry, occupation, salary, and education and skill requirements.

Requisition Number:	FS86446
Interest Category:	Business Operations/Admin/IT
Interest Sub Category:	Administration
Job Title :	Senior Logistics Technician
Employment Category/ Status:	Full-time
Type of Position:	Regular Hire
Country:	U.S.
State:	Indiana
City:	Linton
Minimum Requirements:	• Bachelor's degree from a four-year college or university; or one to two years related experience and/or training; or equivalent combination of education or experience. • Must have computer skills, database knowledge. • Individual must be able to read, analyze, and interpret general business periodicals, professional journals, technical procedures, or governmental regulations.
	...
Job Description:	XXX Corporation is looking for a Senior Logistics Technician to join our team in Linton, Indiana.

Responsibilities:

- Reviews requisitions and negotiates within budgetary limitations and scope of authority.
- Obtains material from supplier at the lowest cost consistent with considerations of quality, reliability of source, and urgency of need.
- Confers with vendors to obtain product or service information such as price, availability, and delivery schedule.

...

XXX Corporation is a leading provider of engineering, construction, and technical services for public agencies and private sector companies around the world. The Company offers a full range of program management; planning, design and engineering; systems engineering and technical assistance; construction and construction management; operations and maintenance; information technology; and decommissioning and closure services. XXX Corporation provides services for power, infrastructure, industrial, oil and gas, and federal projects and programs. Headquartered in San Francisco, XXX Corporation has more than 57,000 employees in a network of offices in nearly 50 countries.

Table 16: Comparison of geographic distributions, UK 2015

	BGT Data	ONS ASHE	Weight
East Midlands	6.23	6.96	1.12
East of England	9.18	9.28	1.01
London	22.69	14.77	0.65
North East	2.09	3.79	1.82
North West	9.24	10.71	1.16
South East	16.01	13.47	0.84
South West	8.09	8.44	1.04
West Midlands	9.83	8.59	0.87
Yorkshire and The Humber	6.7	7.93	1.18
Northern Ireland	1.58	3.05	1.92
Scotland	6.13	8.63	1.41
Wales	2.22	4.38	1.97
Correlation	0.95		

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

Table 17: Occupational distribution of STEM jobs, 2015 (4-digit UK SOC)

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 - 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 18: 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

3 Trend Growth Durations & Shifts

Abstract

Policymakers and investors often conceptualize trend growth as simply a medium/long term average growth rate. In practice, these averages are usually taken over arbitrary periods of time, thereby ignoring the large empirical growth literature which shows that doing so is inappropriate, especially in developing countries where growth is highly unstable. This paper builds on this literature to propose an algorithm, called “iterative Fit and Filter” (iFF), that extracts the trend as a sequence of medium/long term growth averages. iFF separates important country-specific historical episodes and *trend growth durations* - number of years between two consecutive *trend growth shifts*, vary substantially across countries and over time. We relate the conditional probabilities of up and down-shifts in trend growth next year to the country’s current growth environment, level of development, demographics, institutions, economic management and external shocks, and show how both iFF and the predictive model could be employed in practice.

3.1 Introduction

> “The instability of growth rates makes talk of *the* growth rate almost meaningless.”
>
> Lant Pritchett [132]

Many developed and developing countries have been growing at a slower pace since at least the late 2000s. Which ones will continue at a low trend growth rate for another several years? Where could the situation get even worse before it gets any better, and where is a rebound in trend growth imminent?

Such questions are of great interest to many policymakers and investors. However, addressing them in practice is rather challenging for two main reasons. Trend growth is inherently unobserved, with no consensus among economists on how it should be extracted from growth time series data, and a large number of domestic and external factors come into interplay at the very same time, and could either shift the current trend growth rate up, or down, or counterbalance each other and make the country vibrate around the existing trend for another several years.

This paper aims to develop an empirical framework which allows us to address the following two questions: *What is the trend growth rate at which a country is currently* *growing? How likely is it to shift and in what direction?*

Although numerous sophisticated ways of extracting the trend component from growth time series by applying linear, nonlinear, univariate, multivariate filters and other techniques exist (cf., e.g., French [72]), very often, we still want to think about trend growth as merely a medium/long term average growth rate. Despite its simplicity, this latter definition leads to an important practical issue: given a time series of growth rates over *what* periods of time should the averages be taken?

This issue is rarely addressed explicitly, and averages are often taken over subjectively defined periods of time, decades, the whole sample of data that one has access to, etc., thereby ignoring the large empirical growth literature that started in the 1990s with the seminal contributions of Easterly et al. [49] and Pritchett [132], and which shows that employing "arbitrarily chosen long-run average growth rates fails to take account of a very important 'stylised fact' of economic growth, i.e., while the growth process of "developed" economies is well characterized by such a single long run average growth rate (with a "business cycle" around this trend) this is not true of most countries in the world, many of whom exhibit multiple structural breaks in growth rates" (Kar, Pritchett, Raihan, and Sen [107], henceforth KPRS).

Our paper builds on the insights and techniques of this empirical growth literature to first propose a methodology for extracting the trend as a sequence of medium/long term average growth rates taken such that a shift in trend growth, whenever it happens, satisfies economic significance thresholds, and the trend shift dates identified are optimally located in a growth time series. This method, called the "iterative Fit and Filter" (iFF), generalizes the "Fit and Filter" (FF) approach developed in KPRS [107] to identify breaks in growth time series. iFF preserves the merits of the original FF, in particular, the ability to identify a much larger number of breaks in the presence of high growth volatility than purely statistical methods (Bai-Perron [19, 18]), but overcomes the problem of having to postulate a somewhat arbitrary maximum number of breaks in the first step and the fact that, unlike in purely statistical methods, the final break dates identified by the original FF are not necessarily optimally located. A simple example illustrates how both issues matter for whether or not a potential shift date is identified as true.

iFF is therefore a general method which only requires the researcher to specify two parameters: a minimum number of years over which the averages have to be taken (minimum trend growth duration) and the economic significance threshold(s) for the trend growth shifts (trend shifts significance filter). Both parameters should depend on how the researcher defines trend growth for the purpose of his analysis and the type of trend growth dynamics he is interested in understanding and predicting. In this paper, we are interested in medium/long term trend growth and therefore impose a minimum trend growth duration of five years. We also employ the economic significance thresholds proposed in the original FF which have the advantage of recognizing the non-linearity in the growth process by distinguishing between a trend shift in the same direction (e.g. up-shift after up-shift), in which case a 1% point change is enough for significance, and a shift in the opposite direction (e.g. down-shift after up-shift) where the trend has to change by at least 3% points for a significant shift.

Applying iFF with these parameters on a sample of 153 developed and developing countries, we find that it separates important country-specific historical episodes, and that *trend growth durations* - number of years between two consecutive *trend growth shifts* - vary substantially both across countries and for a given country over time. We discuss in some details the experiences of several countries both to illustrate what iFF identifies in practice and to suggest how it could potentially be employed in economic history research where arbitrarily taken averages are still very often the norm.

Despite the heterogeneity of country experiences, several characteristics seem to clearly distinguish the overall trend growth process in developing countries from the one followed by developed economies. For instance, we find that although growth is twice as volatile in developing countries as in developed ones, trend growth is almost three times as volatile and therefore accounts for a larger proportion of the overall growth variance. The probability that we will have to wait for 10 years or more between two consecutive trend growth shifts in a developing country is only 31% versus 72% for a developed economy. Moreover, the median absolute magnitude of a trend growth shift in a developing economy is 5.45% points versus 2.3% points in a developed counterpart. These findings agree with previous research (e.g. Pritchett [132], Aguiar and Gopinath [7]) that documents and investigates the distinction between the smooth and stable growth paths of developed economies and the discontinuous growth patterns in developing countries.

To address our second motivating question - *"How likely is trend growth to shift and in what direction?"* - we model and estimate the conditional probabilities of trend growth up and down-shifts next year, conditional on the country having already grown at the current trend growth rate since the last trend shift. In competing risks models (an extended form of duration/survival analysis with several possible types of events instead of one), these latter objects are known as cause-specific *discrete-time hazard* *rates* (cf., for instance, Allison [11]).

Several empirical growth papers have already employed duration techniques. For instance, Mora and Siotis [127] estimate a discrete-time duration model to analyse how external factors affect recovery prospects in developing economies. Berg et al. [25] and Hausmann et al. [90] employ continuous-time duration analysis to investigate the determinants of the duration of growth spells and growth stagnations respectively. The main advantage of the duration methodology is that it allows researchers to incorporate unfinished/censored episodes into their analysis. In our case, this issue is crucial because all current trend growth episodes are censored. The only thing we know about them is that they have already lasted up until this year. What we do not know and are interested in, is whether the country will continue growing at the current trend growth rate for another several years or experience a trend growth shift next year and in what direction. Duration techniques overcome this problem by focusing on the hazard rate, which, in reality, is just a different representation of the distribution of durations but is unaffected by random censoring (cf. for instance Allignol et al. [10]). The random censoring assumption is satisfied if the fact that the growth episode is censored does not provide additional information for whether the country is more or less likely to experience a trend shift next year, which is the case in our setting since our sample ends in 2015 for all countries.

Historically, duration analysis, especially for competing risks, was first mainly developed in a continuous-time setting (Allison [11]). However, already in his seminal paper, where Cox [42] introduced the partial likelihood method for the estimation of the proportional hazards model, he noted that a continuous-time approach may become problematic if the data contains an "appreciable number of ties" - events recorded as happening at the same time. It is, of course, possible to assume that time is continuous and events are simply "grouped" into discrete time intervals and use approximations in the estimation. In our case, however, we believe that considering time as intrinsically discrete, i.e. "ties [as] real, not spurious" (Grambsch and Therneau [82]), is more appropriate since we are directly working with annual data and a large number of countries experience trend shifts simultaneously.

Moreover, a discrete-time approach has the advantage of giving results that are easier to interpret economically. In discrete time, the hazards are conditional probabilities, whereas in continuous time, they are rates and can therefore be bigger than one. For instance, in our case, working in discrete time leads to modelling the conditional probabilities of trend growth shifting up/down next year, conditional on the current trend growth episode having lasted up to this year. If we were to employ a continuous-time approach, we would be estimating the conditional rates at which trend growth episodes end with up/down-shifts per survival year, something that seems to be less intuitive.

Examining a large set of covariates related to the growth environment, the level of development, demographics, institutions, political stability, economic management, and external shocks, we find, for instance, that while better institutions and higher domestic savings may protect countries from trend down-shifts, higher youth and old dependency ratios are detrimental to trend growth, increasing down-shift risks and hindering up-shifts. Several systemic forces, such as higher average gold and rising food prices (except for food exporters), as well as higher and rising US T-bill rates increase the relative trend growth down-shift risks across the globe. A rise in domestic conflict, credit and inflation may act as catalysts for trend growth down-shifts, whereas devaluations, if not too large, may give a positive impetus to trend growth.

As Pritchett and Summers [133], we also find that "regression to the mean is the empirically most salient feature of economic growth". Whatever the specification, a one percentage point higher trend rate increases the relative down-shift risk and reduces the relative up-shift risk by over 20%, while, on its own, the trend growth rate explains almost 10% of all trend shifts.

We then use the insights from our exploratory analysis to construct a parsimonious model which relates the up and down-shift hazards to 20 different covariates, and estimate the 2016 conditional probabilities of up and down-shifts in trend growth for 120 countries in our sample. These hazards estimates, together with the 2015 estimates of trend growth extracted using iFF for all 153 countries, are contained in Table 25 (cf. Appendix), which constitutes the main output of this paper.

For instance, we find that China has been growing at 6.81% p.a. since 2008 (last trend shift in 2007). 2007 was a trend down-shift, hence a further down-shift would be identified as a $\leq$ 5.81% p.a. average growth rate over the next $\geq$ 5 years, whereas an up-shift would be a $\geq$ 9.81% p.a. average growth rate over the same period. In 2016, the conditional probabilities of trend down and up-shifts are 17.3% and 1% respectively.

The rest of the paper is structured as follows. Section 3.2 presents the trend extraction methodology (iFF) and examples of what it identifies in practice. In Section 3.3, we establish some stylized facts about trend growth durations and shifts in developing versus developed countries. Section 3.4 builds our predictive model by first explaining the econometric framework, then undertaking an exploratory analysis of the potential determinants of trend growth durations and shifts.

3.2 Extracting Trend Growth

Trend growth is unobserved. The assumptions that we impose in order to extract the trend component from growth time series should reflect the way in which we think about the growth process and therefore define what trend growth is. For instance, the well-known Hodrick-Prescott filter [95] extracts trend growth by taking a weighted moving average of the growth time series and therefore assumes that trend growth evolves continuously over time. In this paper, we embrace the basic definition of trend growth as a medium/long term average growth rate and therefore think about the trend as evolving discretely. This section develops a simple iterative algorithm that builds on the "Fit and Filter" (FF) approach proposed in KPRS [107] and extracts trend growth by first identifying economically significant trend shift dates. We start with a precise definition of trend growth in our context, then motivate and describe the trend extraction method, and finally discuss several examples of what it gives in practice.

3.2.1 Defining trend growth

We conceptualize economic growth $\{g(t)\}_{t\in\mathbb{Z}}$ as a process that vibrates around a medium/long term average growth rate - the trend $\{\tau(t)\}_{t\in\mathbb{Z}}$:

$$g(t) = \tau(t) + c(t) \tag{3.1}$$

where the cycle $c(t)$ is a zero-mean transitory fluctuation.

At time t, the trend can shift up:

$$\Delta\tau(t) \equiv \tau(t+1) - \tau(t) > 0$$

or down:

$$\Delta\tau(t) < 0$$

making growth vibrate around a new higher or lower level. Note that $\Delta\tau(t)$ is a forward difference; the trend shift happens in t, but the country starts growing at the new trend growth rate only from period $t+1$.

Suppose that the country is observed for T periods of time over which it experiences $m \geq 0$ trend shifts. As a convention, we set $T_0 = 0$ and $T_{m+1} = T$.[40]

[40] GDP per capita is observed in $[0, T]$, so growth rates can be computed for periods 1 to T.

Let $\mathscr{T} = \{T_1, ..., T_m\}$ be the set of shift dates. We want to extract trend growth as a sequence of medium/long term averages:

$$\tau(t) = \frac{1}{D_j} \sum_{s=T_{j-1}+1}^{T_j} g(s) \text{ for } t \in T_{j-1}+1, ..., T_j, \ j = 1, ..., m+1 \tag{3.2}$$

where $D_j = T_j - T_{j-1}$ is the jth *trend growth duration* - number of years for which the country grows at the trend growth rate at which it started growing in period $T_{j-1} + 1$.

To agree with our basic definition of trend growth as a medium/long term average growth rate that captures some fundamental developments in a country's growth process which go beyond business cycle fluctuations, we want to ensure that:

1. The averages are taken over a *medium/long term*, i.e for all $j = 1, ..., m+1$:

   $$D_j \geq \delta$$

   where δ is the minimum trend growth duration - the least number of periods over which the average has to be taken.

2. Trend shifts are *economically significant,* i.e. for any $t \in \mathscr{T}$:

   $$|\Delta\tau(t)| \geq F$$

   where F is a threshold that we impose. Intuitively, trend shifts have to be large enough, since they should signal some new fundamental developments in the country's growth process.

The choice of both the minimum trend growth duration δ and the threshold(s) F should depend on the type of trend growth movements that we are interested in.

In this paper, we work with annual growth data and think about trend growth as a medium/long term average growth rate, therefore setting $\delta = 5$. If our interest were only in long term growth, we could set, $\delta = 10$ therefore taking averages over at least decade long intervals.

Similarly, if we believe that an at least 2 percentage points change in a medium/long term average growth rate is economically significant, we can set $F = 2$. If our sole interest were in dramatic trend shifts, we could raise F to 5.

The thresholds could also be non-linear. In this paper, we employ the filter proposed in KPRS [107], which sets $F = 2$ for a first shift, then distinguishes between a trend shift in the same direction (e.g. up-shift after up-shift) with $F = 1$, and a trend shift in the opposite direction (e.g. down-shift after up-shift) with $F = 3$. The idea is that once the trend has already shifted up (down), shifting further up (down) by even 1 percentage point is already economically significant. At the same time, in order to avoid confusing trend shifts and business cycles, the medium/long term average would have to change by at least 3 percentage points if shifting in the opposite direction.

There is no single best answer and, as discussed below, no consensus in the academic literature on what the "right" parameters should be. The framework in this paper is therefore designed specifically to be very flexible and general.

In any case, once we have decided on δ and F, all we need to extract trend growth according to eq.3.3 is to identify the trend shift dates $\mathscr{T}$.

3.2.2 Identifying trend shift dates

The statistical and filter approaches Although the ultimate purpose is usually different from extracting trend growth per se, the timing of shifts/breaks in growth time series has been an important preoccupation in many papers. In order to investigate the factors that initiate and halt growth accelerations (Hausmann et al. [89], Berg et al. [25]), growth collapses (Hausmann et al. [90]) or both (Jones and Olken [105], Kerekes [109]), researchers always start by proposing a way of identifying in historical growth data the episodes that are relevant to their study. Since these episodes start with a significant and sustained acceleration/collapse in the average growth rate, their identification relies of the timing of the dates at which such shifts happen.

This empirical growth literature can be broadly classified into two main streams: the papers that use the statistical approach based on the Bai-Perron (BP) methodology [18, 19], and those that employ "filters" - subjectively defined rules that vary from paper to paper.

Given a time series of annual growth rates, the statistical approach (sometimes called the BP methodology) first identifies the sets of break dates that produce the best fit for a given number of breaks, from one up to a maximum. The researcher can impose this maximum number directly and/or specify a minimum number of years between consecutive breaks (akin to the minimum trend growth duration in our case) so that the maximum number of breaks gets determined indirectly by the length of the time series. As mentioned above, there is no consensus on what these numbers should be - Jones & Olken [105] assume a minimum of 5 years between breaks, Berg et al. [25] report results for both 5 and 8, Kerekes [109] opts for 10. In any case, the statistical method then proceeds sequentially: starting from the null hypothesis of no breaks in the time series, it tests whether allowing for additional break(s) significantly improves the goodness of fit.[41] Again, there is no complete agreement on which statistical tests should be employed to gauge this significance - for instance, all previous references [105, 109, 25] employ different tests. This sequential testing continues until we can no longer reject the hypothesis of $m \geq 0$ breaks against the alternative of (one) more break(s) or until reaching the maximum number of breaks allowed/possible.

Whatever the testing procedure, the fundamental problem with the purely statistical approach is that it sometimes identifies *economically* insignificant changes in long term growth averages as *statistically* significant ("false positives"), while omitting some economically significant changes because of statistical insignificance ("true negatives"). For instance, a 2 percentage points change in the medium/long term average growth rate may be identified as statistically significant in a country where the underlying growth process has low volatility so that even small changes seem to be big when viewed through the "statistical" lens. At the same time, a 4 percentage points change can be dismissed in another country where the growth process is inherently more unstable so that important changes appear as "random" from a statistical perspective.

To understand the practical consequences of using the BP methodology for our purpose (extracting trend growth), we implemented a statistical approach based on a standard F-test (Zeileis et al. [164, 163]) and found a statistically significant break for Canada in 1979 with its average real GDP p.c. growth rate changing from 2.7% per annum between 1951 and 1979 to 1.3% p.a. for the 1980-2015 period. We also found only one break for China, in 1977, i.e. we completely missed seven shifts in the Chinese trend growth rate, illustrated in Figure 3.3, which are not only economically large (over 3% points changes each) but, as discussed below, also coincide with major events in the Chinese economic history.

This "low power" issue inherent in the statistical methodology is widely recognized in the literature, e.g. Bai and Perron [20] confirm the presence of "true negatives" through Monte Carlo simulations. Jones and Olken [105] explicitly recognize that the sets of break dates they identify are "conservative", while Berg et al. [25] complement the statistical tests with economic criteria to go from statistical breaks to economically meaningful growth spells by removing some irrelevant breaks.

The "filter" approach avoids this "low power" issue by looking specifically for eco-

[41] The standard practice is to allow for one additional break only. However, sometimes the alternative hypothesis is "one or more" as in Berg et al.[25], i.e. a "double maximum" test, cf. discussion in Kerekes [109].

nomically meaningful changes in the medium/long term growth rate, found by systematically applying a set of researcher-defined rules to growth data. For instance, Hausmann et al. [89] identify growth accelerations as "increases in per-capita growth of 2 percentage points or more, [...] sustained for at least eight years and [such that] the post acceleration growth rate [is] at least 3.5 percent per year." Another example is Hausmann et al. [90], who define growth collapses as "intervals that start with a contraction of output per worker and end when the value immediately preceding the decline is attained again". Clearly, the main disadvantage of the filter approach is the lack of a common framework, which identifies up and down-shifts in a consistent manner.

The original Fit & Filter (FF) A recent paper by Kar, Pritchett, Raihan and Sen [107] summarizes the shortcomings of both approaches and proposes to combine them in order to overcome their limitations while preserving their advantages. The authors call the result "Fit & Filter" (FF) because the approach "involves the best fit of the BP method to the data in the first stage, and the application of a filter to the breaks identified in the first stage in the second stage." Hence, FF overcomes the "low power" of the statistical approach by not using its second step (the statistical tests), while providing a unified way of identifying both economically meaningful up and down-shifts.

Another important advantage of the original FF is that, unlike the standard statistical and filter approaches, it takes into account the nature of the previous shift: a first candidate break is classified as "genuine" if the average growth rate before and after the break changes by at least 2% points. For any subsequent break, the filter distinguishes between a break in the same direction (e.g. acceleration after acceleration), in which case a 1% point change is enough for significance, and a break in the opposite direction (e.g. deceleration after acceleration) where the shift has to be at least 3% points large to be qualified as genuine.

Recognizing this nonlinearity in growth dynamics is important because of the reversion to the mean phenomenon (Easterly et al. [49], Pritchett & Summers [133]) - the idea that it is much easier for countries that have experienced a trend up-shift in the past, to then experience a trend down-shift, i.e. to revert back to the world average growth rate, rather than experience yet another up-shift. Hence a further acceleration in their trend growth rate of even as little as a 1% point is already a substantial achievement. A similar argument would hold for down-shifts.

To help the reader understand how the FF approach works in practice, here is a concrete example from the original paper [107]:

> In the case of Brazil, the first step identifies four candidate break years: 1967, 1980, 1992 and 2002. In 1967, growth accelerated from 3.7% (for 1950–1967) to 6.3% (for 1967–1980). Since this is the first potential break and is above the 2% threshold, we conclude that it as a genuine break. In 1980, growth decelerates from 6.35% to -1.1% (for 1980–1992), a deceleration of 7.4% and easily passes the "deceleration following acceleration" threshold of 3%. In 1992, growth accelerates from -1.1% to 1.4%, a change of 2.5%. However, as this is an acceleration following a deceleration, it would have to be above 3% in order to pass the filter and hence we do not include 1992 as a "genuine" growth break. In 2002, growth accelerated again, this time to 2.5% and since this was an acceleration following a previous candidate acceleration it only had to pass the 1% threshold.

KPRS [107] document that FF achieves a substantial improvement over the statistical method in identifying a larger number of "true negatives", especially in developing countries where the volatility of growth is itself a consequence of a trend growth process with many shifts, and in omitting the "false positives" in developed countries with smooth trend growth paths. Furthermore, the breaks identified often seem to coincide with major events in the economic history of the respective countries. Hence, it seems that FF is fully appropriate as a method of extracting trend growth as a sequence of medium/long term averages taken over periods that are historically meaningful, and such that shifts in trend growth are economically significant.

Despite this, the original FF has two issues which became apparent as we tried to generalize it and employ for our purpose. Both arise because of the way in which FF uses the first step of the BP methodology, i.e. the "Fit" part. As explained above, this first step simply finds the optimal location of a given number of breaks (from one up to a maximum allowed/possible) in a given time series by minimizing the residual sum of squares (best fit). The statistical approach then uses these sets of optimal dates and their associated residual sums of squares sequentially in the second step (the statistical tests). By contrast, there is no sequential testing in the FF: it only uses the optimal set of dates identified for the maximum allowed and hence the choice of this maximum matters a lot.

The authors assume 8 years between breaks and simply "postulate that a country with: (i) Forty years of data, can have a maximum of two breaks. (ii) More than 40 years and up to 55 years, can have a maximum of three breaks. (iii) More than 55 years, can have a maximum of four breaks." However, with five more years of data and assuming a minimum of 5 years between breaks instead of 8, what maximum number should we postulate to apply the filter?[42]

Retaking the example of Brazil. Postulating seven gives 1956, 1962, 1967, 1973, 1980, 1987, and 1992 as potential breakpoints, all of which pass the filter. Choosing eight as the maximum yields 1956, 1962, 1967, 1973, 1980, 1992, 2003, 2010. However, now 1992 and 2010 miss the filter thresholds.

This example illustrates the two crucial issues with the original FF:

1. the choice of the maximum matters for whether we identify a break date as genuine or not: 1992 is a genuine break point if we assume a maximum of seven break dates, but becomes fake if we raise this maximum to eight;

2. the final set of dates identified as genuine by FF is not necessarily optimal: if FF identifies six out of eight breaks as genuine, the locations of the six genuine breaks are not necessarily such that the residual sum of squares is minimized (best fit) over all possible sets of *six* break dates since they were selected as part of the eight-dates set that gives the best fit among all sets of *eight* dates. These are two different optimization problems and the optimal set of m break dates is not necessarily a subset of the optimal set of $m + 1$ break dates (cf. Bai and Perron [19] for the dynamic programming algorithm used to solve these problems). This issue creates a disadvantage for the FF as compared to the statistical/BP methodology where the final set of break points identified is always optimal since only optimal sets are used in the sequential testing and if a set is rejected, *all* the dates within this set are rejected and the non-rejected alternative is just another set of optimally located dates.

The iterative Fit & Filter (iFF) In order to make FF robust to these issues, we propose an iterative algorithm that builds on the original FF and that we therefore call the "iterative Fit & Filter" (iFF). iFF can be easily programmed in any standard statistical software package by following the steps described in the insert on the next page. The computer code that implements it in R is available on request.

[42] We use the Penn World Table version 9.0 (Feenstra et al. [57]) extended to 2015 with IMF World Economic outlook data, while KPRS use PWT version 7.1 that stops in 2010. Also note that we use real GDP p.c. data while KPRS employ GDP p.c. in Purchasing Power Parity.

Algorithm 2 Iterative Fit & Filter (iFF)

Notation: Let $\lfloor x \rfloor$ denote the largest integer that does not exceed x.
For any set of trend shift dates $\mathscr{T} = \{T_1, ..., T_m\}$, the residual sum of squares is computed as:

$$RSS(\mathscr{T}) = \sum_{j=1}^{m+1} \sum_{t=T_{j-1}+1}^{T_j} [g(t) - \tau(t)]^2$$

where:

$$\tau(t) = \frac{1}{D_j} \sum_{s=T_{j-1}+1}^{T_j} g(s) \text{ for } t \in T_{j-1}+1, ..., T_j, \ j = 1, ..., m+1 \tag{3.3}$$

and $D_j = T_j - T_{j-1}$ is the jth *trend growth duration* - number of years for which the country grows at the trend growth rate at which it started growing in period $T_{j-1}+1$.

Step 1: Determine the maximum possible number of trend shifts m_1. Given the length of the time series T and the minimum trend growth duration δ, the growth time series can be divided into at most $\lfloor T/\delta \rfloor$ segments, hence:

$$m_1 = \lfloor T/\delta \rfloor - 1$$

Step 2: Let $\hat{\mathscr{T}} = \{\hat{T}_1, ..., \hat{T}_{m_1}\}$ be the set of m_1 shift dates that minimize the residual sum of squares:

$$\hat{\mathscr{T}} = \arg\min_{\mathscr{T}} RSS(\mathscr{T})$$

over all possible sets of m_1 trend shift dates $\mathscr{T} = \{T_1, ..., T_{m_1}\}$ such that $T_j - T_{j-1} \geq \delta$ for all $j = 1, ..., m_1 + 1$. In practice, this set can be found by using the dynamic programming algorithm described in Bai and Perron [19].

Step 3: Use the optimal trend shift dates $\hat{\mathscr{T}}$ and eq.3.3 to compute the trend $\{\tau(t)\}_{t=1}^{T}$ and the set of trend shifts: $\{\Delta\tau(t)\}_{t \in \hat{\mathscr{T}}}$.

Step 4.1: If all trend shifts satisfy the threshold(s) of the filter, i.e. for all $t \in \hat{\mathscr{T}}$.

$$|\Delta\tau(t)| \geq F$$

we have found an optimally placed set of trend shift dates such that all resulting trend shifts are economically significant. $\{\tau(t)\}_{t=1}^{T}$ computed in Step 3 is the trend.

Step 4.2: While at least one of the trend shifts is not economically significant, we re-iterate Steps 2 and 3 with:

$$m_{k+1} = m_k - 1$$

instead of m_1 until either we end in Step 4.1 with an optimally placed and economically significant set of m_k trend shifts $\hat{\mathscr{T}}$, or $m_{k+1} = 0$ and we conclude that there are no trend shifts and simply compute the trend as the average growth rate over the T periods.

The researcher no longer has to postulate any maximum number of breaks. The only choice parameters are the minimum trend growth duration and the filter, which, as we discussed in section 3.2.1, should depend on how trend growth is defined for the purpose of the research question.

Given the specified minimum duration δ, **Step 1** simply determines the maximum possible number of breaks from the length T of the time series and calls it m_1. We then search for the set of m_1 dates that minimizes the residual sum of squares (**Step 2**) and such that the minimum number of years between any two trend shifts is δ, i.e. we take averages over at least δ years.

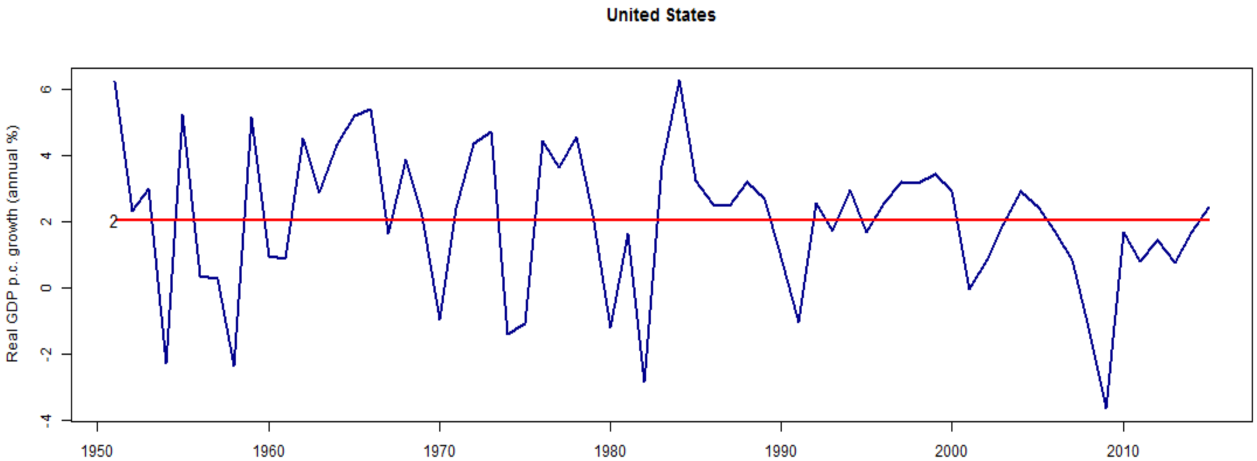
We then use the identified candidate shift dates to compute the candidate trend growth process and the set of candidate trend shifts (**Step 3**). We check whether or not all trend shift dates pass the filter threshold(s).

If yes (**Step 4.1**), we are done: we have found the trend process that satisfies our definition of trend growth - conditions (1) and (2) in section 3.2.1. The optimal trend shift dates delimitate the periods over which the averages have to be taken when extracting the trend from growth time series.

If not (**Step 4.2**), there is at least one trend shift that is not economically significant. To see why we need to re-iterate steps 2 and 3 in this case, suppose that 4 out of 5 trend shifts satisfy the threshold(s). The only thing that we can conclude at this point is that there is no way of segmenting our growth time series with 5 trend shifts that are both optimally placed and economically significant. The four trend shifts that happen to be economically significant are not necessarily optimally placed because their location in step 2 was determined by minimizing the residual sum of squares over all possible sets of *five* trend shift dates. To determine the optimal location of four trend shift dates, we would need to minimize the residual sum of squares over all possible sets of *four* trend shift dates, i.e. re-do step 2 with $m_2 = 4$.

We re-iterate our search as long as a trend process with all trend shifts satisfying the economic filter is not found. Intuitively, in each iteration k, we ask the following question: is it possible to divide our growth time series so that m_k trend shifts are placed optimally (**Step 2**) and are economically significant (**Steps 3 & 4**)? If no trend shift dates are identifed as economically significant, we simply conclude that the country experiences no trend shifts in the sample over which we observe it, and therefore our best guess of its trend growth rate over this sample is the full sample growth average.

Figure 3.1 Trend growth in the United States (1951-2015)



Notes: Trend growth (red) extracted from annual real GDP p.c. growth data (blue) using iFF.

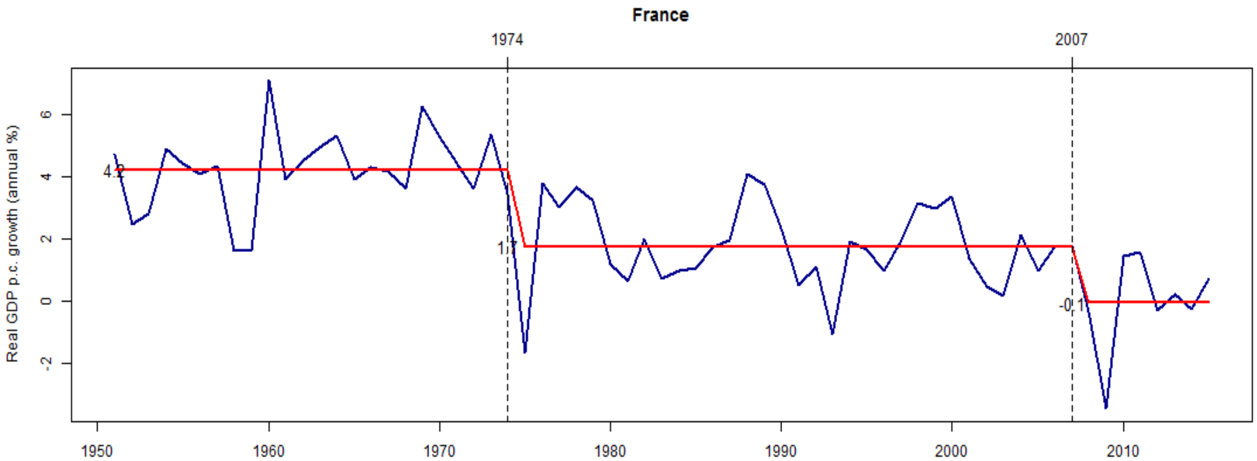
3.2.3 iFF in practice: trend growth & economic history

We now examine some real world examples of what iFF applied to growth time series data yields. Annual real GDP p.c. growth rates are constructed using the *Penn World Table* (PWT) version 9.0 (Feenstra et al. [57]) and extended to 2015 with the IMF *World Economic Outlook* (WEO) data. Our sample contains twenty developed countries and we use developing when referring to any country that is not developed (i.e. our developing countries include newly industrialized countries, emerging markets, frontier markets, and least developed countries).[43]

Table 25 in the Appendix summarizes the most recent (as of 2015) trend growth rate and the last trend shift date and magnitude for all 153 countries in our sample, as well as the hazards (conditional probabilities) of trend up and down-shifts which are the focus of the next section. Figures 3.1 to 3.3 illustrate the complete trend growth paths (red) for the USA, France, and China.

[43]PWT version 9.0 provides data for 182 countries, up to and including 2014. Data goes back to 1950 for some countries. The database is freely accessible at http://www.rug.nl/research/ggdc/data/pwt/pwt-9.0 and fully described in Feenstra et al. [57]. To construct our real GDP p.c. time series, we divide rgdpna (Real GDP at constant 2011 national prices (in mil. 2011US$)) by pop (Population (in millions)). We remove 28 countries with population less than 600,000 in 2014 and the State of Palestine because it is absent from major datasets like the IMF WEO and the World Bank World Development Indicators (WDI) that we employ below. The set of developed countries includes: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom and the United States.

Figure 3.2 Trend growth in France (1951-2015)



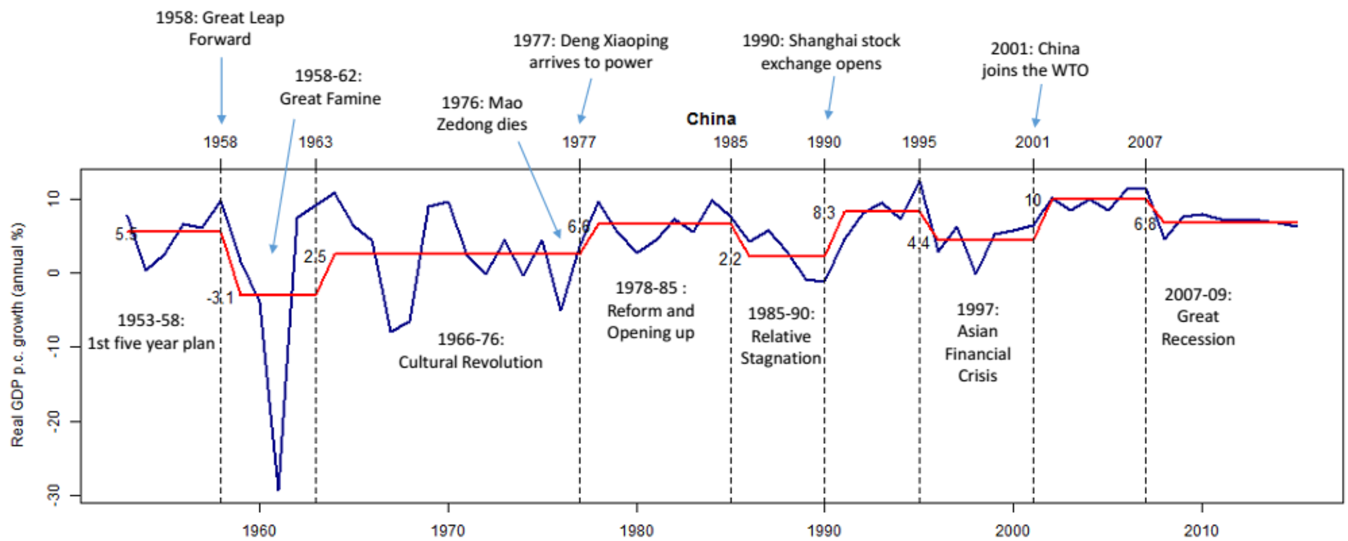
Notes: Trend growth (red) extracted from annual real GDP p.c. growth data (blue) using iFF.

Between 1951 and 2015, trend growth in the USA, as well as Australia, Canada, Sweden and the UK, can be summarized by a single rate of around 2% p.a. At the same time, other developed countries like France, Germany, Italy, Spain etc., and, in particular, many developing economies have experienced more interesting trend growth dynamics that often coincide with important country-specific historical and political developments.

For instance, in France (figure 3.2), 1974/75 marks the end of a thirty years period known as *"The Glorious Thirty"* (*"Les Trente Glorieuses"*, cf. Fourastié [70] who coined the term or Lejeune [116] for a more recent reassessment of the period). Damaged by the two World Wars, France experienced a period of "catch up" growth driven by the reconstruction and the industrialization of the country, rising productivity and consumption levels. However, as this model of growth reached its limits, the country, hit by the 1973 oil shock, entered a period of stagflation, rising unemployment, and slower productivity growth. Trend growth per capita dropped from 4.2% p.a. to 1.7% p.a.

In France, as in Finland, Austria, Ireland, Italy, Greece, Cyprus, Norway and Spain (cf. Table 25), 2007/2008 appears as a down-shift in trend growth, while in other developed economies it is "only" a very large negative fluctuation around a per-existing trend. This may be because this pre-existing trend growth rate was already very low, e.g. Portugal has been growing at close to 0% since 2001, or because the country

Figure 3.3 Trend growth in China (1951-2015)



Notes: Trend growth (red) extracted from annual real GDP p.c. growth data (blue) using iFF.

recovers from the Great Recession relatively quickly and is not as harshly hit by the 2011/12 recession, e.g. Germany.

Note that as time passes and iFF is fit on new enlarged samples, the trend shift dates (especially the latest ones) may be re-assessed. Indeed, with hindsight it will become more obvious whether or not 2007 was a watershed in the economic history of France. As of 2015/16, it seems that the country experienced a downward sustained shift in 2007 not only because of the global financial crisis - although this has certainly been an important catalyst. The profound need for structural reforms in France and a rising level of domestic discontent have rendered the country vulnerable to external shocks. France needs to "change [its] model", to paraphrase the title of a recent book by Aghion, Cette and Cohen [6] in which the authors discuss a set of reforms that could help France become an innovation-driven economy and experience an up-shift from its current 0% trend growth rate.

Another interesting example is China.

After a period of restoration from WWII (1949-1952), the Communist Party of China (CPC), under the leadership of Mao Zedong, launched the first five-year plan in 1953. Modeled after the Soviet example and aided by Soviet planners and engineers, the CPC re-organized industries into cooperatives and farmers into socialized collective units. The main goal of the plan was to achieve high economic growth with a particular focus on developing heavy industries (steel, concrete, iron, machinery, ...). Investment in the industrial sector was financed by extracting surpluses from agriculture where prices were set artificially low. Although, the economy did expand at a trend growth rate of 5.5% p.a., an important sectoral imbalance emerged, and in 1958, the CPC decided to abandon the Soviet model and instead to take a "great leap forward" in the production of all sectors simultaneously.

Collectivization was pushed further with the prohibition of private plots and the establishment of communes. Decision-making and planning were decentralized. The construction of Soviet-like large and capital-intensive plants was pursued but at a slower tempo and now complemented with locally built and run, small-scale, low-technology projects. These "backyard" projects yielded substandard products while diverting an important proportion of farm labour, and together with the inefficiency of the communes, the withdrawal of the Soviet financial and technical support, and several natural disasters, resulted in what is known as the "Great Famine" - a substantial disruption of China's agriculture which starved to death at least 15 million people (unofficial estimates range higher, between 20 and 30 million) between 1959 and 1962.

Indeed, although the CPC started to repel the "Great Leap Forward" program already in 1960 with private plots being returned to the farmers, the communal system being reduced, unemployed workers and investment being transferred from industry to agriculture; it is only in 1963 that the agricultural situation had sufficiently improved and some resources started being redirected back to the industrial sector.

Another consequence of the disruption produced by the "Great Leap Forward" was the appearance of a group of politicians who recognized that China needed to switch to a model of development where material incentives play a greater role, and against whom Mao Zedong initiated the Cultural Revolution in early 1966. Political instability continued until Mao's death in 1976 and Deng Xiaoping's arrival to power in 1977.

Xiaoping announced a modernization program which pushed the country onto a new development path of "reform and opening up" ("*Gaige Kaifang*") from 1978 onwards. Indeed, 1978 is very often considered as a watershed in the economic history of modern China, "the year when China started economic reform" (Zhu [165]). The goal was not to eliminate state planning and control, but to increase the role of material incentives by introducing market mechanisms into the system. The program aimed at expanding foreign trade (by encouraging exports, easing negotiations and cooperation with foreign firms and legalizing trading and credit arrangements) and eliminating existing deficiencies and distortions (e.g. between light and heavy industries).

Zhu [165], who implements a growth accounting exercise to decompose the sources of Chinese growth before and since 1978, shows that in the before period, all growth was due to physical and human capital accumulation, financed by massive government investment and a rise in education levels. On the contrary, since 1978, a rise in productivity became the main driver of growth.

Productivity in agriculture was stimulated through a substantial rise in official agriculture prices and the creation of the "household responsibility system" in 1979. Under the latter, farmers had to sell a certain amount of grains at official prices to the government, but could then transact anything beyond the quota at market prices and for their own enrichment.

The results were considerable: fig. 3.3 shows that the economy expanded at 6.6% p.a. and per capita. Trade increased from 8.5% of GDP in 1977 to 23% in 1985. All sectors were expanding except for manufacturing, whose value added as a percentage of GDP even fell from 39.3% in 1977 to 34% in 1985, because heavy industry was purposefully restrained.[44]

The reforms were introduced gradually, first in a few localities, then, if successful, nationally, and completed by 1984 when most households were under the responsibility system, and most communes had been dissolved. The efficiency gains "from workers using the same technology with a much more rewarding set of incentives, were largely exhausted" by 1984 (Zhu [165]), and a period of relative stagnation began with trend growth shifting down to 2.2% p.a. in 1985.

A new liberalization wave emerged around 1990 with trend growth shifting up to 8.3% p.a. Government interventions were further reduced, markets for agricultural inputs and outputs further liberalized and incentives for the adoption of new technologies set in place. The Shanghai and Shenzhen stock exchanges opened in 1990 and 1991 respectively. The "open door" policy was introduced in 1992 with the aim of creating a legal basis for Chinese-foreign joint ventures. Special economic zones were introduced to facilitate the influx of foreign investment.

Although China was less affected by the Asian Financial crisis than other economies, e.g. Indonesia and Thailand which experienced trend down-shifts to negative growth rates in 1996, the Chinese trend growth rate did slow down to 4.4% p.a. in 1995. Indeed, foreign direct investment as a percentage of GDP fell from 6% in 1994 to 3.2% in 2000 and monetary conditions were tightened with the annual broad money growth rate falling from 31.5% in 1994 to 12.3% in 2000.

The last up-shift in China's trend growth rate to 10% p.a. in 2001 coincides not

[44]Figures quoted in this passage are taken from the World Bank Development Indicators.

only with a global economic recovery but also China's entry into the World Trade Organization, which "introduced international economic laws [into the country] and ushered a period of rapid regulatory reform by creating agencies such as the China Banking Regulatory Commission (CBRC) to govern increasingly globally integrated markets" (Yueh [162]).

Recently, there has been a lot of debate in the press and the academic literature (e.g. Pritchett & Summers [133]) about whether or not the Chinese economy is slowing down. Our analysis indicates that China, hit by several natural disasters (the 2008 Chinese winter storms and floods in the South, the Sichuan earthquake) and the Global Recession (e.g. Chinese trade fell from 64.8% of GDP in 2006 to 41.2% in 2015), has already experienced a down-shift in trend growth in 2007 and is currently growing at 6.8% p.a.

Whether a further slowdown will happen in the future is an open question and we shall try to contribute to the ongoing debate below, in section 3.4.3.

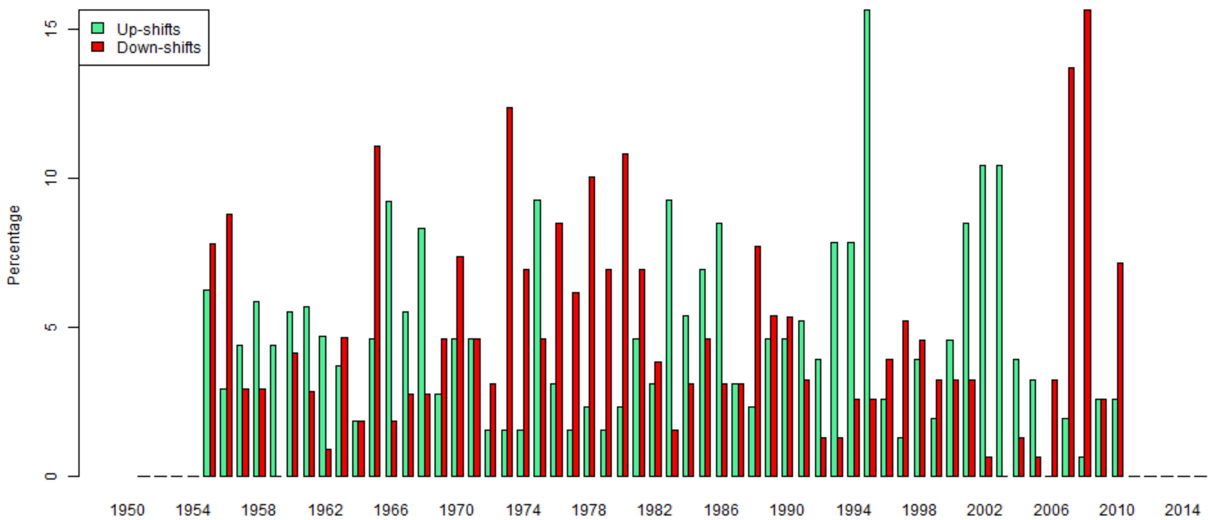
These examples suggest that iFF inherits from the original FF the ability to identify in a systematic way important episodes in the economic history of a country and could therefore be a useful tool for economic historians, many of whom still often rely on their judgment or arbitrary time periods when presenting and interpreting summary statistics. For instance, to decompose the 1978-2007 period into subperiods, Zhu [165] simply takes three ten years long periods: 1978-1988, 1988-1998, and 1998-2007, while our analysis suggests a rather different decomposition of the 1978-2007 period into economically and historically meaningful subperiods. In future work, iFF could therefore be employed to undertake a much more thorough economic history analysis for a larger number of countries.

3.3 Trend Growth Durations & Shifts: Stylized Facts

Although the previous examples illustrate the great variety of trend growth dynamics within and between countries, it is also important to try to establish some stylized facts about the trend growth processes identified. In this section, we look at the temporal and spatial distributions of *trend growth shifts*, and the distribution of *trend growth durations* in developed versus developing countries.

3.3.1 Trend growth shifts

Figure 3.4 Temporal distribution of trend shifts



Note: Percentages of countries experiencing up and down-shifts in trend growth p.a.

Temporal distribution Although we have growth data for 153 different countries, these countries are observed over different periods of time between 1950 and 2015, so that the effective number of countries in our sample varies over time.[45]

Hence, the number of trend shifts happening in a given year is not directly comparable over time. Instead, in order to investigate the temporal distribution of trend shifts, Figure 3.4 shows the percentage of countries experiencing trend up and down-shifts each year.

The earliest and latest shifts happen respectively in 1955 and 2010 since we assume a minimum trend growth duration of 5 years. The alternation between red bars dominating green ones and vice versa suggests that up and down-shifts do not coincide. Indeed, the correlation between the percentages of up and down-shifts each year is -0.12; not very big but negative, possibly indicating the presence of systemic trend growth spillovers that could be investigated in more details in future work.

Earlier studies (e.g. Ben-David & Papell [23]) have found that for developed coun-

[45]For instance, data for all post-Soviet nations only starts in 1990 (growth data from 1991). These countries have so far been almost always excluded from similar studies on the grounds of not sufficiently long time series - e.g. Hausmann et al. [89] require at least 20 data points, Pritchett & Summers [133] exclude all countries with "less than 25 years of data". We believe it is important to integrate this group of countries into our study, not only because we now do have 25 years of growth data for them, but also because their post-Soviet experiences certainly contain a lot of valuable information for helping us understand what affects trend growth durations & shifts.

tries most up-shifts take place in the 1950s and the 1960s during the postwar reconstruction period, which coincides with a significant liberalization of trade and the creation of institutions such as Bretton-Woods and GATT, while most down-shifts take place in the early 1970s, with the first oil price shock in 1973 acting as an important catalyst.

Our research confirms and extends these previous findings. All twenty developed countries are observed between 1950 and 2015 and, overall, experience 11 up-shifts and 31 down-shifts. Most up-shifts (8 out of 11) happen during the 1950s and 1960s, and 35% of all down-shifts take place in the 1970s. However, a novel finding is that 32% of all down-shifts happen in the 2000s highlighting the magnitude of the impact of the Great Recession in the developed world.

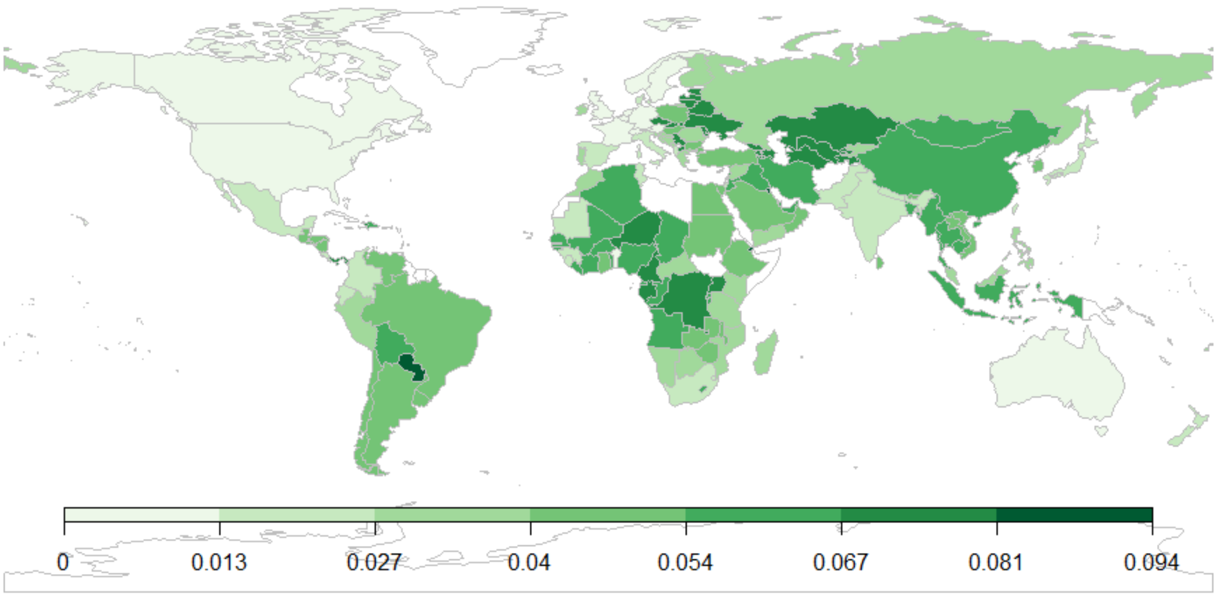
Many Latin American countries experience trend down-shifts in the late 1970s/early 1980s, e.g. Venezuela in 1977, Argentina and Brazil in 1980, Mexico and Chile in 1981 (cf. Diaz-Alejandro [47] for a thorough analysis of the Latin American debt crisis). Most down-shifts of the late 1990s are related to the Asian Financial crisis: Thailand and Indonesia in 1996, Malaysia and Singapore in 1997...

The 15% of up-shifts in 1995 mostly come from countries that were part of the Soviet Bloc - Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Uzbekistan, Serbia, Slovakia, Slovenia, Macedonia ... which started to recover from the disruptions in their economies produced by the collapse of the Soviet Union in 1991. Another important wave of up-shifts happens in the early 2000s. Some are further up-shifts in emerging eastern Europe (Tajikistan & Ukraine in 2001, Armenia in 2000...), others are recoveries from the Asian crisis (Indonesia and Thailand 2001). Many commodity exporters experience up-shifts in 2002/2003 (Argentina, Peru, and Colombia in 2002, Bolivia and Venezuela in 2003) as a recovery from a period of historically low commodity prices between 1998 and 2002 which led to significant falls in tax revenues and important economic disruptions in these countries (cf. Tenreyro [152] for a retrospective analysis of the 2001-2002 Argentine crisis and Spatafora and Samake [151] for an empirical investigation of how commodity prices and fiscal outcomes are related).

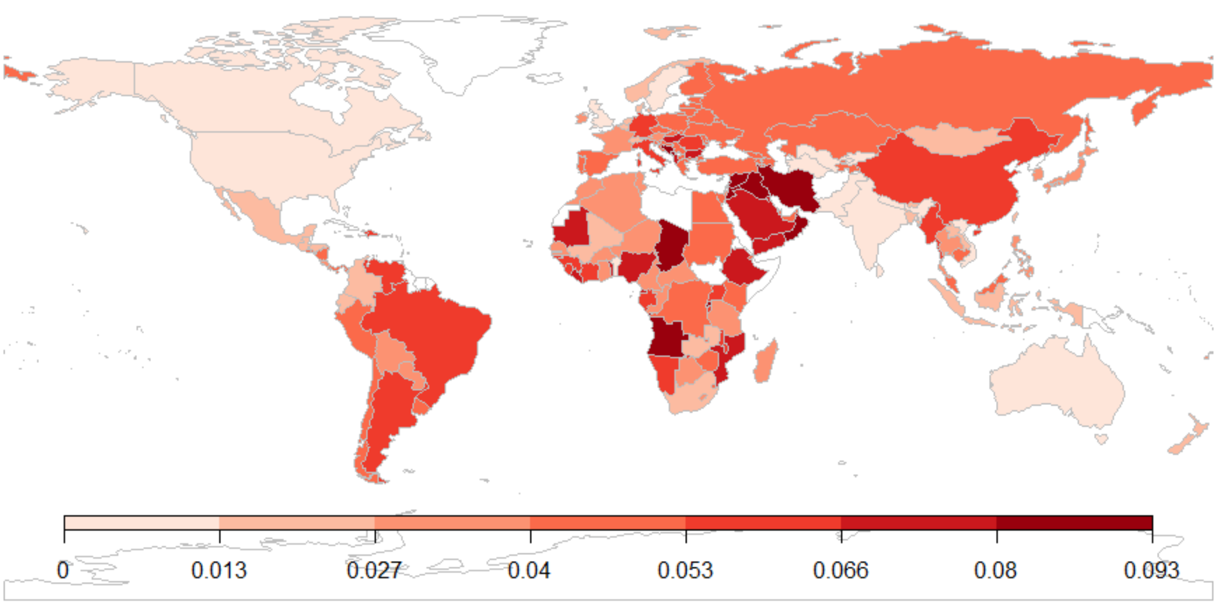
The late 2000s, especially 2007/2008, stand out as the most important years of trend growth down-shifts in modern history. Of course, we should not forget that this is close to the end of the sample and results could change later on, with hindsight. However, this finding seems plausible given the magnitude of the growth disruptions provoked by the Great Recession in both developed and developing countries.

Figure 3.5 Spatial distribution of trend growth shifts

Up-shifts per annum



Down-shifts per annum



Note: Countries not in sample are in white.

Spatial distribution Each country in our sample has between 25 and 65 years of growth data. Hence, similarly to the number of trend shifts per annum, the numbers of trend shifts per country are not directly comparable. Instead, to investigate the spatial distribution of trend up and down-shifts, Figure 3.5 illustrates the respective numbers of shifts per annum.

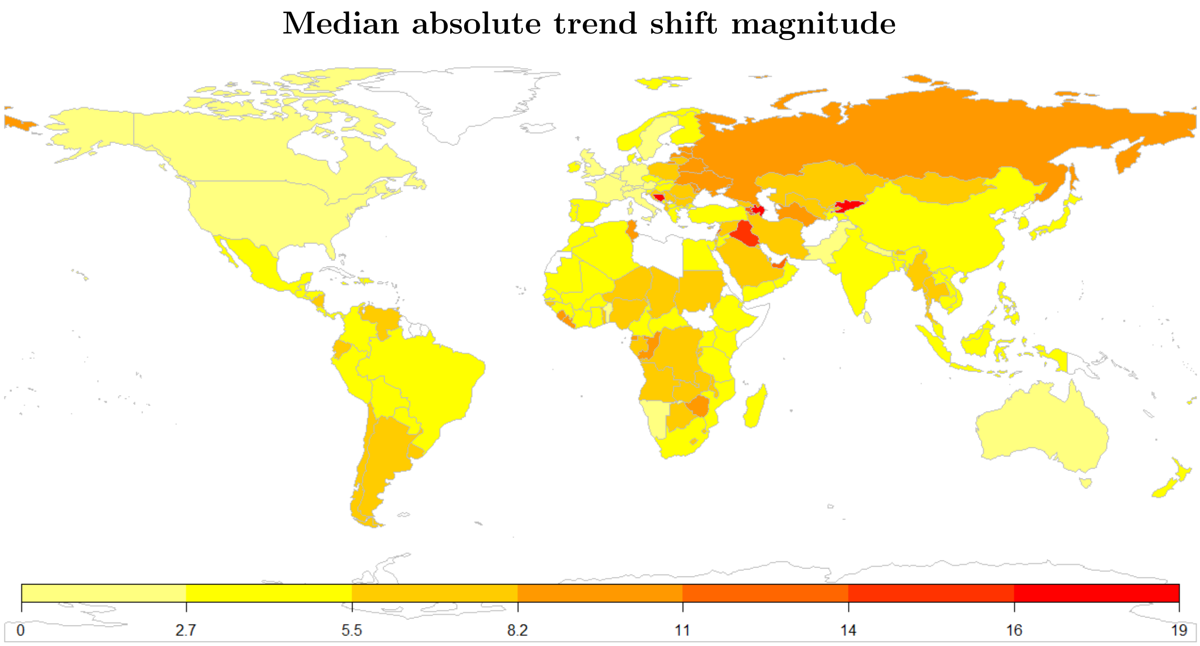
Several interesting observations emerge. First of all, developing countries are more prone to both up and down-shifts than developed ones. This finding is not new. In his seminal contribution, which spurred researchers to pay much more attention to within country growth dynamics, Pritchett [132] already argued that "a single time trend does not adequately characterize the evolution of GDP per capita in most developing countries".

The correlation between the numbers of up and down-shifts p.a. is 0.33, suggesting that for some countries trend growth is unstable in both directions, and that, at least to some extent, all countries are capable of both up and down-shifts. In particular, the figures buttress the findings of Frances, Paap and van Dijk [71] who examine the question of whether Africa is less capable of growth than Latin America or Asia. They implement a data-based classification of countries into clusters and find that one third of African countries are not assigned to the low growth cluster. Hence, it is wrong to aggregate and simply label Africa as the "lost continent". Figure 3.5 indeed shows that although some African countries do exhibit very large numbers of down-shifts p.a., this is not the case for all the continent. Moreover, many African countries also exhibit a significant number of up-shifts and are comparable to Latin American and Asian countries in terms of numbers of up and down-shifts. If anything, it seems that there is more heterogeneity in both up and down-shifts p.a. in Africa than on any other continent.

Middle East countries appear as very prone to down-shifts because of the numerous conflicts that have taken place in these countries over the past half century.

One potential caveat to bear in mind when interpreting our findings, is that some of our up-shifts are from negative growth to less negative growth, while some down-shifts are from positive to less positive growth. In this simple approach, we also do not distinguish between shifts of different magnitudes. Figure 3.6 therefore complements the analysis by showing the spatial distribution of the median absolute trend shift magnitude for each country in our sample. Once again, it seems that overall developing nations experience much larger swings in their trend growth paths in both directions. The median absolute trend shift in a developing country is 5.45% points against a mere

Figure 3.6 Spatial distribution of trend shift magnitudes



Note: Countries not in sample are in white.

2.3% pts. for developed economies.

Note that the relatively large trend shift magnitudes of many post-Soviet nations may be directly due to the fact that we have growth data for them only since 1991, a period that coincides with a particularly turbulent part of their histories after the collapse of the Soviet Union. Their large up-shift intensities exhibited in Figure 3.5 are also, at least to some extent, the result of recovering from a period of very negative growth rates in the first half of the 1990s.

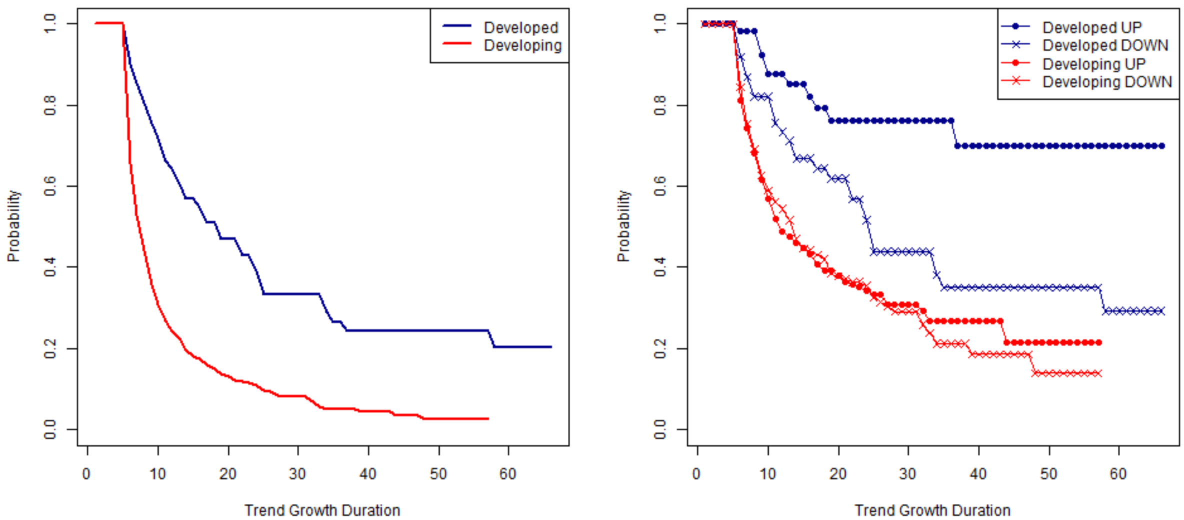
3.3.2 Trend growth durations

In practice, the much greater instability of trend growth in developing countries documented above implies that the number of years over which it makes economic and historical sense to take medium/long term averages of growth rates - *trend growth durations* - in these countries are shorter.

To make this more precise, let's think of the *trend growth duration* - number of years between two consecutive *trend growth shifts* - as a discrete random variable that takes values in $\{\delta, \delta+1, \delta+2, ...\}$.

Our data consists of trend growth episodes, each starting in $T_{j-1}+1$ (the year after the last trend growth shift or the beginning of the sample if $T_{j-1} = 0$, where 0 is the

Figure 3.7 Trend growth survivor functions



Notes: Developing countries trend growth survivor functions estimation based on 671 trend growth episodes and 629 trend growth shifts. For developed countries the respective numbers are 62 and 42.

year when GDP p.c. data is first observed for the country so that growth rates are computed from year 1) and ending in T_j with trend growth shifting either up or down, or with the end of the sample if $T_j = T$. Hence, each trend growth episode has a certain duration $D_j = T_j - T_{j-1}$. When $T_j = T$, the duration is censored since we do not know when the current trend growth episodes will end. The only thing we know is that they have already lasted for $\widetilde{D}_m = T - T_{m-1}$ years, m being the number of trend growth shifts.

Treating our data as a random sample collected from the population of trend growth episodes, the left panel of Figure 3.7 compares the probabilities that an developing (red) versus a developed (blue) country has a trend growth duration of at least d years:

$$S(d) = \Pr(D_j \geq d) = \sum_{k=d}^{\infty} \Pr(D_j = k)$$

In survival/duration analysis, $S(d)$ is known as the *survivor function.*[46]

The comparison is stark: while the probability of having 10 years or more between two consecutive trend shifts in a developed country is 0.72, in a developing country it is as low as 0.31. In developed countries, half of the trend growth episodes last for at least

[46]Many great reference on discrete time survival analysis exist, for instance, Allison [11], Rodriguez [135], or Jenkins [100].

18 years ($S(18) = 0.5$), while in developing countries this number is around 7 years.

Also, note that 20% of all trend growth episodes from developed countries in our sample are censored at 66 because a number of developed countries experience no trend shifts between 1950 and 2015, i.e. we only know that these countries have been growing at the same constant trend growth rate for at least 66 years over which we observe them.

Since we have assumed $\delta = 5$, the probability of all trend growth durations lasting at least 5 years is one: $\Pr(D \geq 5) = 1$.

Thereafter, $S(d)$ is computed as:

$$S(d) = \prod_{k=\delta+1}^{d-1} (1 - \alpha(k)) \tag{3.4}$$

where $\alpha(d)$ is the *discrete-time hazard* - the conditional probability that a trend growth shift happens exactly d years after the last one, conditional on the current trend growth episode having already lasted d years:

$$\alpha(d) = \Pr(T_j = T_{j-1} + d | D_j \geq d) = \frac{\Pr(D_j = d)}{S(d)} \tag{3.5}$$

The intuition behind eq.3.4 is that for a trend growth episode to last for, say, at least eight years, the country must "survive" at the same trend growth rate for at least five years. This is always true since $\Pr(D \geq 5) = 1$. Conditional on this, it must survive the sixth year without trend shifts. This happens with probability $(1 - \alpha(6))$. Given all this, if no shift happens in the seventh year - an event that has probability $(1 - \alpha(7))$, the trend growth episode will have lasted 8 years or more: $S(8)$.[47]

Since, any trend growth episode can end with either a trend up-shift or a trend down-shift, the overall discrete hazard can be decomposed into two cause-specific hazards:

$$\alpha(d) = \alpha^U(d) + \alpha^D(d)$$

[47] The non-parametric estimates shown in Figure 3.7 are constructed as follows (discrete version of the Kaplan-Meier non-parametric estimator):

$$\hat{S}(d) = \prod_{k=\delta+1}^{d-1} (1 - \frac{r_k}{n_k})$$

where r_k is the number of trend growth episodes lasting for k years, while n_k is the number of episodes that could potentially last k years, i.e. are "at risk" of ending k years after beginning. In survival literature, $\frac{r_k}{n_k}$ is often called the "exit rate".

where

$$\alpha^U(d) = \Pr(T_j = T_{j-1} + d, \Delta\tau(T_j) > 0 | D_j \geq d) \tag{3.6}$$

$$\alpha^D(d) = \Pr(T_j = T_{j-1} + d, \Delta\tau(T_j) < 0 | D_j \geq d) \tag{3.7}$$

The right panel of Figure 3.7 divides the survivor functions into up and down-shift specific curves:

$$S^U(d) = \prod_{k=\delta+1}^{d-1} (1 - \alpha^U(k)) \tag{3.8}$$

and similarly for $S^D(d)$.

Interestingly, while in developing countries, one has to wait for about the same number of years before an up or a down-shift in trend growth occurs; in developed economies, we have to be much more patient before up-shifts than before down-shifts. Concretely, the probability that we will have to wait 10 years or more until observing a trend up-shift in a developing country is 0.57. This number for a down-shift is very close: 0.59. By contrast, in developed countries the respective numbers are 0.88 and 0.82 and, as illustrated in Figure 3.7, the difference between the two probabilities grows very quickly as we consider longer horizons. For instance, the probability that we will have to wait 20 years or more before observing an up-shift in a developed economy is still very high 0.76 and well above the probability of waiting the same time until a down-shift: 0.62.

The intuition behind these findings can be related to standard economic growth theories as follows. In the process of convergence towards developed-economy status, countries experiences a similar number of up and down-shifts. Once developed and located at the technology frontier, however, generating up-shifts becomes much more difficult: "catch-up" growth is no longer available and growth spurts have to either come from exogenous technical progress (neo-classical growth model, Solow [150]), or an increase in savings as a proportion of GDP (AK model, Romer [136]), or innovation (Schumpeterian model, Aghion and Howitt [5]). At the same time, down-shifts can still happen because of exogenous shocks, or as a direct consequence of growth slowing down due to convergence (think about the post WWII reconstruction period slowdowns in most European countries).

Finally, Table 19 summarizes some stylized facts from this section about how the growth processes differ in developing and developed economies.

Although, overall, growth is twice as volatile in developing countries as in devel-

Table 19: Trend growth in developing vs. developed countries

	Developing	*Developed*
Std. Growth*	5.3	2.5
Std. Trend Growth*	3.4	1.2
Trend Var. as % of Growth Var.*	41.4	29.5
Up-shifts p.a.*	0.046	0
Down-shifts p.a.*	0.044	0.015
Median absolute shift magnitude*	5.45	2.30
Pr(10 years or more between shifts)	0.310	0.719
Pr(10 years or more until up-shift)	0.568	0.877
Pr(10 years or more until down-shift)	0.590	0.821
Number of Countries	133	20
Number of Up-shifts	322	11
Number of Down-shifts	307	31

Notes: *Median across countries. Std. (standard deviations) in percentage points.

oped ones, the trend is almost three times more volatile and therefore accounts for over 41% of the total growth variance against slightly less than 30% in developed economies. This corroborates recent research by Aguiar and Gopinath [7] which shows that "emerging markets are characterized by a volatile trend that determines the behavior of the economy at business cycle frequencies". The result implies that understanding the determinants of the trend growth dynamics in developing countries is more important than in developed economies even in the medium term.

Several conclusions emerge from assessing the overall characteristics of trend growth durations and shifts. The trend growth path is much more unstable in developing economies as a group and represents a higher proportion of the overall growth process. Developing countries experience more trend shifts with larger trend shift magnitudes. However, contrary to developed countries, where waiting for up-shifts takes longer, up and down-shifts in developing economies happen roughly at the same rate. Despite these stylized facts, figures 3.5 and 3.6 illustrate the vast heterogeneity in country experiences within groups, while the country-specific examples discussed earlier (subsection 3.2.3) suggest that an important amount of variation in the duration of trend growth episodes exists for a given country over time. This implies that a discrete divide into developing/developed is too simplistic and calls for a more systematic assessment of the domestic and external macroeconomic factors that make trend growth durations vary so much within and between countries.

3.4 Predicting Trend Growth Shifts

Section 3.2 developed a methodology (iFF) for determining the trend growth rate at which a country is currently growing by extracting the trend from growth time series as a sequence of medium/long term growth averages taken over economically and historically meaningful periods of time. The purpose of the present section is to address our second motivating question: *"How likely is trend growth to shift and in what direction?"*

3.4.1 Econometric framework: a trend-shifting model of growth

We start by extending the conceptual framework introduced in section 3.2.1 to achieve a comprehensive description of trend growth dynamics.

Remember that we conceptualize growth $g(t)$ as a process that vibrates around a trend:

$$g(t) = \tau(t) + c(t) \tag{3.9}$$

The cycle $c(t)$ is a zero-mean transitory fluctuation and the trend $\tau(t)$ is defined as the average growth rate between two consecutive trend growth shifts:

$$\tau(t) = \frac{1}{D_j} \sum_{s=T_{j-1}+1}^{T_j} g(s) \text{ for } t \in T_{j-1}+1, ..., T_j,\ j = 1, ..., m+1 \tag{3.10}$$

where $D_j = T_j - T_{j-1}$ is the duration of the trend growth episode that starts in $T_{j-1}+1$.

The trend is subject to competing domestic and external forces, summarized in the vector $\mathbf{x}_{t-1}$, which, if strong enough, can shift the trend up or down at time t and make growth vibrate around a new higher or lower level from $t+1$ onwards. In order to rule out a reverse effect from the trend shift in year t to the level of the time-varying variables in year t, we only use time $t-1$ information to predict time t trend shifts.

In particular, our goal is to model the conditional probabilities of trend growth shifting up or down at time t, given that the last trend shift (or the beginning of the sample) happened d years ago :

$$\Pr(T_j = t, \Delta\tau(T_j) > 0 | T_j - T_{j-1} \geq d, \mathbf{x}_{t-1}) \tag{3.11}$$

$$\Pr(T_j = t, \Delta\tau(T_j) < 0 | T_j - T_{j-1} \geq d, \mathbf{x}_{t-1}) \tag{3.12}$$

as functions of the competing forces $\mathbf{x}_{t-1}$.

Without the conditioning on $\mathbf{x}_{t-1}$, equations 3.11 and 3.12 are nothing else than the up and down-shift hazard rates $\alpha^U(d)$ and $\alpha^D(d)$, i.e. equations 3.6 and 3.7, rewritten in calendar time t instead of duration time d by noting that $d = t - T_{j-1}$ and $D_j = T_j - T_{j-1}$.

This arises because we think of trend shifts as recurrent events and adopt what in the survival literature is sometimes called a "reset-clock" approach. After each trend shift, we reset the clock to zero, and once the minimum trend growth duration has elapsed, the country becomes once again "at risk" of experiencing yet another trend shift.

An important assumption underlying "reset-clock" specifications is that "the processes affecting the occurrence of the first event are the same as those for the second, third, and later events" (Allison [11]). In our case, however, trying to disentangle the competing forces that have systematically been important determinants of trend growth durations and shifts is precisely the goal, hence this assumption seems fully appropriate.

A potentially more important limitation of a "reset-clock" specification is the assumption that the hazards do not depend on all the event history, i.e. only on when the last trend shift happened but not when the previous shifts had happened. However, this assumption can be easily relaxed by introducing explanatory variables that represent the dependency of the hazard on the country's previous history (Allison [11]), which is what we do by including among our covariates variables like the trend growth rate and growth volatility, estimated on rolling samples, i.e. from the beginning of the sample up to and including time $t-1$, thereby taking into account all the previous growth history of the country.

Another issue in models with repeatable events is the intra-subject correlation arising from having multiple observations (and potentially also multiple events) per country. In what follows, we adopt a so-called "marginal approach" (cf. Grambsch and Therneau [82] (chapter 8)), which does not include country random or fixed effects, but corrects the standard variance estimates for intra-country correlations. Different approaches could be explored in future work.

In reality, we have already started investigating the determinants of trend growth durations in the previous section where we examined how survivor functions change depending on whether the country is a developing or a developed economy. Since we were only interested in the effect of one specific characteristic - an indicator for being a developing economy, we could proceed in a simple, intuitive way: divide our sample into developed and developing countries, construct the survivor functions non-parametrically as explained in footnote 47, plot the results and inspect them visually.

Unfortunately, this simple approach does not work if our goal is to examine the simultaneous effects of several discrete and continuous characteristics on the hazards. Moreover, it does not give us one quantitative statistic that summarizes the effect of a characteristic and which would allow us to gauge its statistical and economic significance, and to compare it to the effects of other characteristics.

A simple way around these issues, is to assume a specific functional form that relates the hazards to the characteristics $\mathbf{x}_{t-1}$. Since the discrete-time hazards are conditional probabilities, the functional form needs to be such that the estimated hazards lie between 0 and 1 and the hazards of the three possible outcomes - up-shift, down-shift and no shift - sum to one.

In Cox's [42] original paper, where he proposed the partial likelihood method for the estimation of the proportional hazards model for continuous-time survival analysis, he also suggested that a logit specification could be employed in the discrete case and reduces to a proportional hazards model when the time interval considered gets very small. Later, the model was extended to the competing risks situation by relating the covariates $\mathbf{x}_{t-1}$ to the hazards through a multinomial logit specification (e.g. Allison [11], Allignol et al. [10]):

$$\alpha^S(t) = \frac{\exp(\mathbf{x}_{t-1}'\beta^S)}{1 + \sum_{S=U,D} \exp(\mathbf{x}_{t-1}'\beta^S)} \tag{3.13}$$

where, for ease of notation, we write $\alpha^S(t) \coloneqq \alpha^S(t - T_{j-1}|\mathbf{x}_{t-1})$ with $S = U, D$. The parameters β^S capture the cause-specific effects of the covariates on the S outcome relative to no shift.

To see the intuition behind this functional form, suppose a country reaches year t without having yet experienced a novel shift in its trend growth since the last one, and the minimum trend growth duration has elapsed. What can happen in year t? The trend can either shift up, shift down or not shift at all. The problem is therefore akin to a conditional multinomial choice model where the conditional probabilities of the three possible events/choices are:

$$\alpha^U(t), \alpha^D(t) \text{ and } 1-\alpha(t)$$

Note that equation 3.13 indeed ensures that the estimated conditional probabilities lie between 0 and 1, and sum up to 1, since the conditional probability of no shift (reference category) is:

$$1-\alpha(t) = \frac{1}{1+\sum_{S=U,D}\exp(\mathbf{x}'_{t-1}\beta^S)} \tag{3.14}$$

Appendix 3.6.1 explains in details how the likelihood function in our case should be constructed and therefore how the model parameters can be estimated by maximum likelihood.

To see how the parameters should be interpreted, suppose we have two covariates: x_{1t-1} and x_{2t-1}. The probability of experiencing an up-shift in year t relative to experiencing no-shift in year t, conditional on having experienced no shifts since T_{j-1} is:

$$\frac{\alpha^U(t)}{1-\alpha(t)} = \exp(\beta_1^U x_{1t-1} + \beta_2^U x_{2t-1}) \tag{3.15}$$

$\exp(\beta_1^U x_{1t-1} + \beta_2^U x_{2t-1})$ is often called the relative risk associated with covariate values x_{1t-1} and x_{2t-1} . More precisely, $\exp(\beta_1^U x_{1t-1} + \beta_2^U x_{2t-1})$ is the risk of an up-shift relative to no shift. Taking logs:

$$\log\left(\frac{\alpha^U(t)}{1-\alpha(t)}\right) = \beta_1^U x_{1t-1} + \beta_2^U x_{2t-1} \tag{3.16}$$

hence β_1^U measures the change in the multinomial log-odds of an up-shift in trend growth relative to no shift due to a one unit change in x_{1t-1} holding x_{2t-1} fixed.

Another, perhaps more intuitive, interpretation arises from writing:

$$\frac{\exp(\beta_1^U(x_{1t-1}+1) + \beta_2^U x_{2t-1})}{\exp(\beta_1^U x_{1t-1} + \beta_2^U x_{2t-1})} \quad = \quad \exp(\beta_1^U) \tag{3.17}$$

i.e. when x_{1t-1} increases by one unit while x_{2t-1} is fixed, the relative up-shift risk is multiplied by $\exp(\beta_1^U)$. For instance if $\exp(\beta_1^U) = 0.8$, the risk of an up-shift relative to no-shift falls by 20%. A value of 1 means there is no effect on the relative up-shift risk. If $\exp(\beta_1^U) > 1$, the risk rises by $(\exp(\beta_1^U) - 1)$%.

3.4.2 Results: time-to-shift determinants

The economic history examples discussed above point to two types of variables that could be of potential interest:

- some characteristics of the growth, political, institutional, external ... environment which either create favorable (detrimental) conditions for trend up-shifts, or protect the country from (make it more vulnerable to) down-shifts;
- certain shocks which act as catalysts.

There is an important trade-off in selecting variables for such a large heterogeneous set of countries. On the one hand, we want to be able to estimate the model on a relatively large sample so that our results are not driven by a selected few experiences. On the other hand, not including certain variables may lead to an omitted variable bias and affect our coefficient estimates and significance.

In what follows, we examine a large set of covariates which can be regrouped into five dimensions:

1. *Growth environment:* we investigate the effects of the trend growth rate, the number of years since the last shift, the cyclical component and growth volatility (standard deviation of annual growth rates). All these variables are estimated on rolling samples, i.e. for instance, we only use growth data up to and including t to estimate the trend in year t. Using rolling samples is important because the estimates of trend growth may change as we fit iFF on enlarged samples if different shift dates are identified. Hence, using trend growth estimated on the complete sample would be forward-looking. A similar argument can be advanced for the other three variables. For instance, using a time-invariant estimate of growth volatility based on the whole sample of data introduces a look-ahead bias as it implies that at any point in time we know what would be happening to growth in the future. It is true that using rolling samples, especially when these are relatively small, introduces measurement error and could attenuate our estimated coefficients. However, given our interest in employing this model as a predictive tool, the look-ahead bias issue seems more important.

2. *Development and Demography:* instead of a discrete classification of countries into "developed" and "developing", we employ a set of variables which capture both material and non-material aspects of development: the real GDP p.c. in

Purchasing Power Parity[48], fertility, infant mortality, life expectancy, the level of urbanisation, primary, secondary and tertiary gross enrollment ratios, the percentage of total population aged less than 14 and the percentage of total population aged 65 and above. We examine these variables both in levels and changes over the past 2 years.

3. *Institutions and political stability:* we use the POLITY database described in Gurr et al. [85]. In particular, the Polity 2 score, measured on a scale from -10 (strongly autocratic) to +10 (strongly democratic), changes in the Polity 2 score over the past two years, and the durability of the regime's authority (number of years since the last substantive change in authority characteristics defined as a 3-point change in the Polity score, cf. [85]). To capture political stability, we employ the Cross-National Time-Series Data Archive (CNTS). In particular, the Weighted Conflict Index, which is a weighted sum of the number of assassinations, general strikes, guerrilla warfare, government crises, purges, riots, revolutions and anti-government demonstrations from the Domestic Conflict Event Data part of the CNTS. The compilation methods and the construction of the index are explained in Wilson [160]. We look at the level and growth in WCI. The latter is winsorized at 100% and a dummy variable equal to 1 when the WCI increases by 100% or more is included. We also use several variables from the Political Data part of CNTS: the number of Coups d'Etats, Major Constitutional Changes, Major Cabinet Changes, Changes in Effective Executive, and the number of Legislative Elections.

4. *Economic management*: we consider the annual inflation rate (GDP Deflator), domestic credit to the private sector (% of GDP), gross capital formation (% of GDP), gross domestic savings (% of GDP). We examine these variables both in levels and percentage point changes. Two variables capture trade: exports plus imports (as % of GDP), and the difference between the annual growth rates of imports and exports. We also look at the annual depreciation of the official nominal exchange rate against the US dollar. The variable is winsorized above at 100%, and a dummy tracks winsorized observations. We allow the effect of the depreciation to be different in the case of a fixed exchange rate regime by including a term which interacts the depreciation with a dummy equal to one if

[48] Real GDP per capita in PPP is constructed using the PWT (rgdpe/pop) and extended to 2015 with World Bank data. All remaining variables are taken from the World Bank Development Indicators, unless another data source is explicitly specified.

the exchange rate is fixed.[49]

5. *External environment & shocks:* we take the annual average and the annual percentage point change in the daily US T-bill (secondary market, 3 month rate). For commodity prices, we use the IFS monthly gold, food and oil price indices. We consider the annual averages, and the growth rates between January and December. We also interact the annual growth in food/oil prices with a dummy equal to 1 if the country is a food/oil exporter and another dummy equal to 1 if the country is a food/oil importer. The overall impact of the annual growth in food/oil prices in year $t-1$ therefore enters the hazard functions as follows:

$$(\beta + \beta_X \mathrm{I}^X_{t-1} + \beta_M \mathrm{I}^M_{t-1}) x_{t-1}$$

   where I^X_{t-1} is a dummy equal to 1 if the food/fuel exports represent at least 20% of merchandise exports in year $t-1$. Similarly, I^M_{t-1} is an indicator function that takes a value of 1 in $t-1$ if food/fuel imports represent at least 20% of merchandise imports in that year.

Given the trade-off between sample size and omitted variable bias, we proceed sequentially, examining one/two additional categories of variables at a time while keeping those that have been previously identified as significant. We fit the models both with and without five-year dummies to check whether the effects estimated are robust to the inclusion of some time-varying unobserved heterogeneity. We do not use a full set of year dummies because it leads to significant over-fitting and non-convergence of the likelihood function, especially in more complex specifications. Moreover, in our exploratory exercise, unlike the predictive model of the following subsection, we shall use the same sets of covariates for both up and down-shift hazards since, a priori, we do not know which covariates enter which hazard function.

Note that all result tables show exponentiated coefficients, i.e. the interpretation is in terms of relative risk ratios, as explained above.

We start with a baseline specification which only includes the *Growth environment* variables: Table 20, models (1) and (2).

The most significant variable in economic and statistical terms is the *Trend growth rate*. Considering the first specification, a one percentage point higher trend growth rate reduces the relative up-shift risk and increases the relative down-shift risk by about

[49] We use the IMF AREAER database for the classification of exchange rate regimes.

Table 20: Growth environment, development and demography

	(1)		(2)		(3)		(4)	
	Up	*Down*	*Up*	*Down*	*Up*	*Down*	*Up*	*Down*
Trend growth rate	0.789***	1.201***	0.774***	1.204***	0.763***	1.237***	0.774***	1.243***
	(-9.85)	(7.11)	(-9.90)	(7.93)	(-7.01)	(5.14)	(-6.85)	(6.29)
Years since last shift	0.925***	0.959***	0.927***	0.957***	0.934***	0.959***	0.937***	0.958***
	(-6.40)	(-4.25)	(-5.69)	(-4.08)	(-4.52)	(-3.80)	(-4.34)	(-3.72)
Cycle	0.934***	1.125***	0.932***	1.112***	0.905***	1.156***	0.905***	1.147***
	(-4.11)	(5.81)	(-4.31)	(4.82)	(-4.80)	(5.18)	(-4.66)	(4.75)
Growth Volatility	1.056***	1.042	1.061***	1.040	1.027	1.019	1.026	1.020
	(2.71)	(1.55)	(2.91)	(1.30)	(1.44)	(0.86)	(1.40)	(0.81)
Log(Real GDP p.c. PPP)					0.477***	1.383*	0.473***	1.351*
					(-3.32)	(1.88)	(-3.06)	(1.68)
Fertility (births per woman)					0.883	1.689***	0.940	1.621***
					(-0.70)	(4.04)	(-0.34)	(3.76)
Infant Mortality					0.989	1.002	0.987	1.000
					(-1.38)	(0.22)	(-1.60)	(0.00)
Life Expectancy					0.966	0.974	0.952	0.965
					(-1.04)	(-0.98)	(-1.50)	(-1.27)
Urban Population (% total)					1.001	1.005	1.002	1.006
					(0.06)	(0.75)	(0.16)	(0.93)
Fertility (change)					0.670	1.167	0.719	1.003
					(-1.13)	(0.35)	(-0.86)	(0.01)
Infant Mortality (change)					1.008	1.015	1.017	1.015
					(0.32)	(0.70)	(0.64)	(0.71)
Life Expectancy (change)					1.096	0.993	1.151**	0.960
					(1.54)	(-0.10)	(2.24)	(-0.60)
Urban Population (ppt. change)					0.959	1.046	0.968	1.020
					(-0.67)	(0.85)	(-0.53)	(0.40)
Primary Enrol. (% gross)					0.999	1.008*	1.001	1.007
					(-0.19)	(1.73)	(0.18)	(1.61)
Secondary Enrol. (% gross)					1.003	0.997	1.003	0.998
					(0.44)	(-0.49)	(0.38)	(-0.29)
Tertiary Enrol. (% gross)					1.002	1.007	1.001	1.008
					(0.17)	(0.89)	(0.12)	(0.99)
Primary Enrol. (ppt. change)					0.988	0.994	0.983	0.999
					(-0.86)	(-0.53)	(-1.23)	(-0.10)
Secondary Enrol. (ppt. change)					1.016	1.016	1.014	1.014
					(0.98)	(1.39)	(0.79)	(1.31)
Tertiary Enrol. (ppt. change)					1.057*	1.010	1.038	1.021
					(1.88)	(0.37)	(1.18)	(0.83)
Pop. ages 0-14 (% of total)					0.934*	0.938*	0.917**	0.943*
					(-1.76)	(-1.90)	(-2.17)	(-1.76)
Pop. ages >=65 (% of total)					0.857***	0.971	0.841***	0.965
					(-2.86)	(-0.68)	(-3.10)	(-0.80)
Pop. ages 0-14 (ppt. change)					1.069	1.004	1.042	1.028
					(0.66)	(0.04)	(0.39)	(0.27)
Pop. ages >=65 (ppt. change)					1.497	1.146	1.488	1.235
					(1.25)	(0.55)	(1.24)	(0.84)
Five Year dummies	No		Yes		No		Yes	
Observations	4062		4062		2856		2856	
Pseudo R^2	0.129		0.152		0.180		0.204	

Notes: Exponentiated coefficients; t statistics in parentheses. Standard errors clustered at country level.
Changes taken over past two years. All variables lagged one year.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

20%. This result illustrates the regression to the mean phenomenon - the idea that an extended period of high growth is rarely sustainable for a long time and more likely to be followed by a period of average rather than even higher growth. Pritchett and Summers [133] have extensively studied this phenomenon in a recent paper, concluding that empirically it is the "most salient feature of economic growth". Our study confirms this finding since trend growth remains the most significant variable throughout the analysis. On its own (regression not shown), the trend growth rate explains almost 10% of trend up and down-shifts.

The duration of the current trend growth episode (*Years since last shift*) has a negative impact on both up and down-shift relative risks. Intuitively, if the country has been growing at the same trend growth rate for a longer period of time, it is more likely that this rate corresponds to a long run equilibrium from which the country is less likely to be destabilized. This variable might also be capturing to some extent the level of development since, as shown in Figure 3.7, trend growth durations are longer in developed economies. Indeed, as we introduce variables from our development bucket into the regression, the magnitude of the effect of duration on up-shifts drops from one additional year since the last trend shift reducing the relative up-shift risk by 7.5% (model 1) to decreasing it by 6.6% (model 3).

The effect of the *Cycle* goes in the same direction as that of the trend: decreases the likelihood of up-shifts and increases that of down-shifts. Mechanically, the reason can be explained as follows: the cycle is computed as growth minus the rolling estimate of the trend. Hence, a higher cycle indicates that our rolling estimate of trend growth is going up and that a trend down-shift from this higher estimate is even more likely. Therefore, the cycle accentuates the effect of the trend that we have just discussed.

Higher *Growth volatility* increases both up and down-shift relative risks although the effect is statistically insignificant for down-shifts and becomes also insignificant for up-shifts as development variables are introduced. Economically, the effect goes in the expected direction though - countries with higher growth volatility have shorter trend growth episodes - hence, we decided to keep this variable as a control.

Including five-year dummies in specification (2) does not greatly affect the magnitude and significance of the growth environment variables, suggesting that these variables are robust to some unobserved time-varying heterogeneity.

Models (3) and (4) incorporate *Development and Demography* variables.

A higher level of *Real GDP p.c. in PPP* decreases the likelihood of up-shifts and increases that of down-shifts. This result agrees with our earlier discussion of economic convergence at the end of section 3.3.2: as countries become more developed, growth slows down. Relative to the no shift outcome, countries start experiencing less up-shifts and more down-shifts. The larger and more significant effect on the up-shifts was expected from our previous discussion and the right panel of Figure 3.7. The coefficients are more difficult to interpret this time since we are looking at the log of the real GDP p.c. The unexponentiated coefficients are -0.74 and 0.32; hence a 10% higher real GDP p.c. this year is associated with a 7% fall in the up-shift relative risk and a 3% rise in the down-shift relative risk next year.[50]

Real GDP p.c. only captures the material aspect of development. For instance, in 2015, Qatar, Singapore, and the United Arab Emirates ranked above the developed country with the highest real GDP p.c. - Norway. Hence, the simple developed/developing divide gets blurred when we think about development in continuous terms. Unfortunately, very few variables reflecting the non-material side of development appear to be significant. This is not surprising given the results from previous studies. For instance, Berg et al. [25] examine the significance of primary and secondary education, adult and child mortality (levels and within growth spell changes) for the duration of growth spells. They only find three out of eight variables to be significant at the 10% level.

A high level of *Fertility* appears to be detrimental to trend growth: one extra child per woman increases the down-shift relative risk by over 60%. Historically, very high fertility rates (7-8 children per woman) characterize several African countries, e.g. Rwanda, Kenya, Oman, Jordan, in the 1970s and 1980s. The damaging effect of fertility could arise because high fertility is often the flipside of a lower level of female education and employment which are detrimental to growth and that we are not controlling for in the regression because this type of data is less common and less reliable.

Higher percentages of *Population aged 0 to 14* or *65 and above* reduce the relative risk of up-shifts, with the effect being bigger and more significant for the 65 and above age bracket. The effect is consistent with larger dependency ratios preventing the savings rate to rise and engender a trend growth up-shift as in a standard AK growth model (Romer [136]). Conditional on fertility, *Population aged 0 to 14* also appears to protect from down-shifts to some extent, perhaps indicating a potential positive effect

[50]To see where these numbers come from note that: $\exp(\beta \ln x_1)/\exp(\beta \ln x_2) = \exp(\beta \ln(\frac{x_1}{x_2}))$. Hence, for instance, a 10% higher real GDP p.c. multiplies the up-shift relative risk by $\exp(-0.74 \ln(1.1)) \simeq 0.93$, a 7% decrease.

on trend growth arising from a future younger and larger active labour force.[51]

An argument often put forward for the case that China still has decades left to run at high growth rates before slowing down is that its level of urbanization, which currently stands at 55.6% of the total population, is much lower than that of the USA (81.6%). Given the results in Table 20, we can neither buttress nor reject this argument because none of the variables related to urbanization are significant. However, it seems that the effect of an increase in *Urban population*, although insignificant, goes in the direction of increasing the likelihood of down-shifts while decreasing that of up-shifts.

A one percentage point increase in the *Tertiary education* gross enrollment ratio over the past two years raises the up-shift relative risk by 5.7% suggesting, as expected, that more higher education is beneficial for trend growth. The small positive effect of *Primary education* on the down-shift relative risk is less intuitive. Both education effects are only significant at the 10% level and not robust to the inclusion of five-year dummies in specification (4). On the other hand, the beneficial effect of an improvement in *Life expectancy*, only appears as significant once time effects are included. This weakness in robustness to the inclusion of other covariates is confirmed in the next set of results, shown in Table 21, where all three variables completely loose significance once we control for the quality of *Institutions, political stability* and *Economic management*.

Acemoglu et al.[1] provide an extensive overview of the various channels through which weaker institutions may disrupt long term growth. In agreement with this, we find that a one unit lower *Polity 2 score* (less democratic institutions) increases the down-shift relative risk by around 5%. Changes in the Polity score are also very important: as expected, a one unit *Amelioration* over the past two years reduces the down-shift relative risk, while a one unit *Deterioration* increases it. Interestingly, the Polity score and changes thereof have no significant effect on the up-shifts, perhaps indicating that good institutions on their own are not enough to substantially lift trend growth. The *Durability of the Polity regime* reduces both up and down-shifts suggesting that political stability leads to growth stability.

Only one out of the six variables from the CNTS database examined happens to be significant: a dummy equal to 1 if the Weighted Conflict Index (WCI) rises by more than 100%. *Conflict rise >=100%* multiplies the relative risk of down-shift by 2. The finding that the remaining political stability variables are insignificant seems a bit surprising. We investigated whether this may be due to the fact that it takes more

[51]See Higgins [94] for an investigation of how demography and national savings are related and a discussion of the dependency debate.

Table 21: Institutions, political stability, and economic management

	(5)		(6)		(7)		(8)	
	Up	*Down*	*Up*	*Down*	*Up*	*Down*	*Up*	*Down*
Trend growth rate	0.747***	1.388***	0.756***	1.376***	0.787***	1.372***	0.778***	1.379***
	(-6.72)	(6.58)	(-6.70)	(6.48)	(-6.80)	(7.24)	(-7.38)	(7.28)
Years since last shift	0.947***	0.981	0.951***	0.983	0.939***	0.986	0.943***	0.986
	(-2.91)	(-1.50)	(-2.68)	(-1.30)	(-3.55)	(-1.22)	(-3.18)	(-1.22)
Cycle	0.891***	1.235***	0.887***	1.218***	0.926***	1.208***	0.919***	1.198***
	(-3.59)	(4.47)	(-3.62)	(4.18)	(-2.78)	(5.95)	(-2.96)	(5.35)
Growth Volatility	1.025	0.998	1.025	1.001	1.017	1.012	1.023	1.011
	(1.28)	(-0.10)	(1.26)	(0.03)	(0.85)	(0.79)	(1.11)	(0.70)
Log(Real GDP p.c. PPP)	0.527***	1.560*	0.501***	1.481	0.553***	1.687***	0.527***	1.604**
	(-2.80)	(1.77)	(-3.02)	(1.47)	(-3.78)	(2.79)	(-3.84)	(2.44)
Fertility (births per woman)	0.710	1.873***	0.777	1.749**	0.801	1.380**	0.762*	1.406*
	(-1.62)	(2.83)	(-1.16)	(2.36)	(-1.48)	(2.03)	(-1.65)	(1.81)
Pop. ages 0-14 (% total)	0.956	0.952	0.923**	0.957	0.944**	0.997	0.941*	0.994
	(-1.17)	(-0.89)	(-1.99)	(-0.76)	(-2.05)	(-0.07)	(-1.93)	(-0.12)
Pop. ages >=65 (% total)	0.939	1.022	0.898*	1.026	0.932*	1.058	0.913*	1.066
	(-1.09)	(0.30)	(-1.74)	(0.32)	(-1.67)	(0.99)	(-1.95)	(1.05)
Primary Enrol. (% gross)	0.997	1.008	0.997	1.007				
	(-0.44)	(1.41)	(-0.38)	(1.16)				
Tertiary Enrol. (ppt. change)	1.047	1.025	1.027	1.043				
	(1.18)	(1.03)	(0.62)	(1.61)				
Life Expectancy (change)	1.088	0.982	1.105	0.973				
	(1.00)	(-0.28)	(1.14)	(-0.41)				
Polity 2 score (level)	0.993	0.955**	0.999	0.953**	0.973	0.947***	0.976	0.948***
	(-0.30)	(-2.27)	(-0.04)	(-2.27)	(-1.48)	(-3.24)	(-1.26)	(-2.85)
Amelioration Polity 2	1.007	0.667***	0.995	0.677***	0.980	0.858***	0.984	0.872**
	(0.17)	(-3.12)	(-0.12)	(-3.04)	(-0.52)	(-2.69)	(-0.42)	(-2.25)
Deterioration Polity 2	0.933	1.147**	0.911	1.164**	0.997	1.157***	0.988	1.151**
	(-1.15)	(2.19)	(-1.48)	(2.22)	(-0.07)	(2.60)	(-0.28)	(2.53)
Durability Polity regime	0.983***	0.982***	0.982***	0.981***	0.977***	0.980***	0.977***	0.980***
	(-2.92)	(-3.69)	(-2.99)	(-3.67)	(-3.62)	(-4.05)	(-3.56)	(-3.98)
Coups d'Etats	0.433	2.177	0.415	2.572				
	(-0.76)	(0.88)	(-0.81)	(1.06)				
Major Constitutional Changes	0.853	1.660	0.910	1.529				
	(-0.39)	(1.25)	(-0.22)	(1.02)				
Changes in Effective Executive	1.048	0.963	1.076	0.946				
	(0.18)	(-0.10)	(0.28)	(-0.15)				
Legislative Election	0.998	0.836	0.998	0.856				
	(-0.01)	(-0.68)	(-0.01)	(-0.57)				
Weighted Conflict Index (WCI)	1.000	1.000	1.000	1.000				
	(0.25)	(-0.61)	(0.69)	(-0.73)				
WCI growth, win.	1.001	1.001	1.000	1.001				
	(0.69)	(0.43)	(0.20)	(0.74)				
Conflict rise >=100%	0.931	2.420***	0.992	2.260**	0.874	1.963**	0.867	1.942**
	(-0.18)	(2.70)	(-0.02)	(2.39)	(-0.43)	(2.39)	(-0.44)	(2.33)
Log(1+inflation)	1.441	3.959***	1.392	3.123**	1.259	3.011***	1.235	2.558**
	(0.82)	(2.71)	(0.79)	(2.39)	(0.60)	(2.70)	(0.56)	(2.48)
Capital formation (gross, % GDP)	0.983	0.991	0.982	0.995				
	(-1.31)	(-0.66)	(-1.27)	(-0.35)				
Capital formation (ppt. change)	0.996	0.995	0.999	0.994				
	(-0.13)	(-0.16)	(-0.05)	(-0.19)				
Domestic savings (gross, % of GDP)	1.002	1.004	1.004	1.003				
	(0.30)	(0.48)	(0.56)	(0.36)				
Domestic savings (ppt. change)	1.038*	0.958*	1.038*	0.970	1.022	0.959***	1.024	0.970**
	(1.96)	(-1.76)	(1.91)	(-1.21)	(1.28)	(-2.92)	(1.24)	(-2.02)
Domestic credit to private (% GDP)	0.997	1.004	0.996	1.003				
	(-0.59)	(1.12)	(-0.68)	(0.87)				
Domestic credit to private (ppt. change)	0.996	1.026***	0.996	1.030***	0.996	1.033***	0.998	1.036***
	(-0.33)	(2.78)	(-0.30)	(2.61)	(-0.37)	(3.81)	(-0.19)	(3.55)

	(5) Up	(5) Down	(6) Up	(6) Down	(7) Up	(7) Down	(8) Up	(8) Down
Depreciation (LCU/$), win.	1.015**	0.976*	1.014**	0.983	1.014***	0.987	1.013***	0.991
	(2.56)	(-1.78)	(2.39)	(-1.35)	(3.06)	(-1.58)	(2.85)	(-1.13)
Depreciation x Fixed Exchange	0.998	1.019	1.000	1.015				
	(-0.23)	(1.28)	(-0.06)	(1.04)				
Depreciation >=100%	0.0673***	0.443	0.0802***	0.446	0.156***	0.405	0.182**	0.402
	(-2.88)	(-0.51)	(-2.87)	(-0.56)	(-2.61)	(-0.60)	(-2.48)	(-0.71)
Imports gr. - Exports gr.	1.004	1.018**	1.006	1.017**	0.999	1.016***	1.001	1.016***
	(0.50)	(2.52)	(0.73)	(2.46)	(-0.22)	(2.90)	(0.08)	(2.99)
Trade (% of GDP)	1.010***	1.001	1.012***	1.001	1.003*	1.000	1.004**	1.000
	(2.98)	(0.44)	(3.38)	(0.38)	(1.86)	(0.21)	(2.44)	(0.29)
Five Year dummies	No		Yes		No		Yes	
Observations	2194		2194		2625		2625	
Pseudo R^2	0.243		0.264		0.213		0.237	

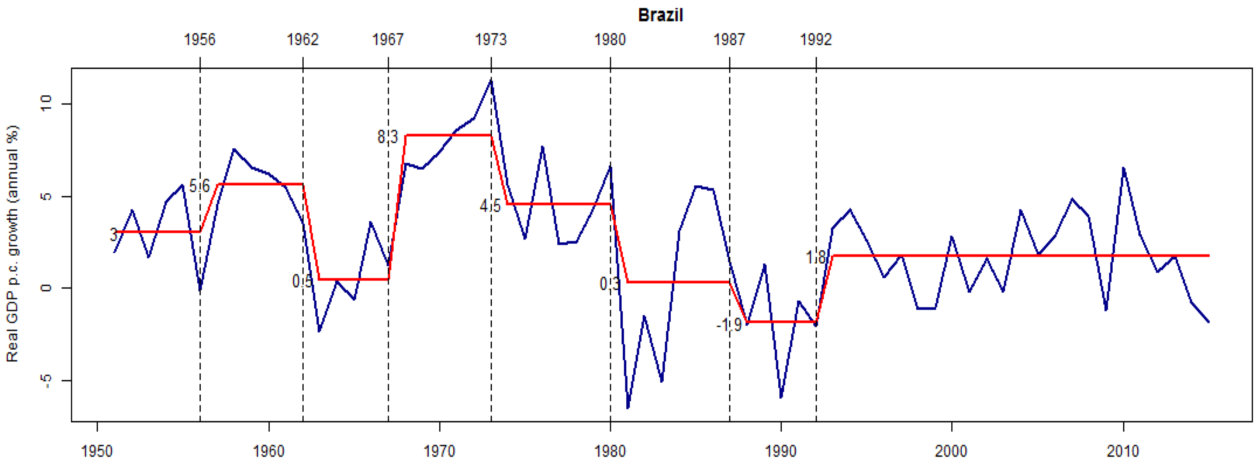
Exponentiated coefficients; t statistics in parentheses. Standard errors clustered at country level. All variables lagged one year.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

than one year for the event to have an impact on trend growth by re-fitting the model with the same variables either lagged two years or aggregated over the past two years, however, no more significant effects appeared. We also investigated the components of the WCI separately, but again with no success. Perhaps, this result is at least in part due to the quality of the data. The CNTS derives most of the events used to construct the WCI from the New York Times (cf. Wilson [160]), and it is very likely that many events, especially in the developing world, go unrecorded. It is also worth mentioning, however, that previous studies have also found the effect of conflict not to be robust to the inclusion of other covariates. For instance, Hausmann et al. [90] find that an indicator for war becomes insignificant for the probability of growth collapses when variables like inflation and the change in the Polity score are included in the regression.

We examine several aspects of *Economic management* starting with monetary stability as measured by the log of 1 plus the annual *Inflation* rate, a standard transformation in the empirical growth literature. Higher inflation significantly increases the likelihood of down-shifts. This result is not surprising; many examples in economic history indicate that inflation is a symptom of economic mismanagement. For instance, all four trend growth down-shifts in Brazil, illustrated in Figure 3.8, were preceded by very high inflation rates: 31% in 1961, 19% in 1972, 56% in 1979 and 145% in 1986, while the success of the *"Plano Real"* (1994-2002), which managed to stabilize the Brazilian economy at the current trend growth rate of 1.8% p.a. since 1993, relied to a large extent on having achieved monetary stability through measures like a peg of the Brazilian real to the US dollar and a general indexation of prices (e.g. cf. Feijo et al. [58]).

Another important determinant of macroeconomic stability is the exchange rate. We look at the *Depreciation* of the official nominal exchange rate against the US dollar. The

Figure 3.8 Trend growth in Brazil (1951-2015)



Notes: Trend growth (red) extracted from annual real GDP p.c. growth data (blue) using iFF.

coefficient estimates indicate that a one percentage point larger *Depreciation* raises the up-shift relative risk by about 1.5%, and this effect does not disappear if the exchange rate is fixed. One particular channel through which a devaluation this year can boost growth next year is by making the country's exports more competitive. However, the dummy indicator for a devaluation of 100% and more (*Depreciation>=100%*) almost completely annihilates the up-shift relative risk.

Just as human capital accumulation, physical *Capital formation* does not seem to be a significant predictor of trend growth dynamics. This agrees with two studies that employ growth accounting techniques (parametric in Jones and Olken [105], non-parametric in Kerekes [109]) to investigate the sources of important growth changes and conclude that factor accumulation only plays a negligible role in them so that "even medium-run growth rate changes are mainly the result of productivity changes"[109].

Two other variables that matter for trend shifts are domestic indebtedness and savings. Both variables are significant in percentage point changes but not in levels. The positive effect of a rise in *Domestic savings* on up-shifts disappears as other insignificant covariates are dropped from the up-shift hazard function, cf. specifications (7) and (8). However, the protective effect from down-shifts remains significant: a one ppt. increase in domestic savings this year reduces the relative down-shift risk next year by around 3%. In developing economies, higher internal savings may help avoid the dramatic economic disruptions caused by sudden outflows of foreign investment. In developed economies, increases in domestic savings can help finance innovation and therefore prevent growth from shifting down.

Increases in domestic indebtedness played a prominent role in the most recent down-shifts. However, as the following table suggests, the lesson that a rise in *Domestic credit to the private sector* may not be good for trend growth could perhaps have been assimilated already after the Latin American turmoil in the 1980s and the Asian crisis of the 1990s.

Country	Down-shift	Rise in domestic credit to private sector in preceding year (ppt. of GDP)
Spain	2007	20.6
Ireland	2007	19.7
Cyprus	2008	18.0
Malaysia	1997	17.2
Thailand	1996	13.1
Chile	1981	10.6

We also find that while a higher proportion of GDP in *Trade* is prospicious to trend growth up-shifts, a rising trade deficit (*Imports growth - Exports growth)* is a symptom of future trend down-shifts.

The remaining specifications in Table 21 show that adding five-year dummies and dropping insignificant covariates does not qualitatively alter the effects just identified in most cases, even though the precise quantitative estimates may change.

Table 22 keeps the variables previously identified as significant and adds the last category: *External environment & shocks.*

The *US T-bill rate* is considered as the risk-free rate on the market and is therefore an important determinant of borrowing costs. Model (9) indicates that both high US rates and rises thereof are bad for trend growth. For instance, a one percentage point higher US T-bill rate is associated with an 8.5% higher relative down-shift risk and an 8.2% lower relative up-shift risk. Interestingly, as we include five year dummies in specification (10), the significance of the US T-bill rate drops. The only effect that remains significant is that of the average US T-bill rate on the relative down-shift risk, which is now much bigger: 17% instead of the previous 8.5%. Although US rate hikes are not significant for trend shifts once we control for unobserved time-varying heterogeneity, their significance in the absence of such controls indicates that they may be a good proxy for a part of this unobserved heterogeneity which is detrimental for long-run growth, in particular for up-shifts.

Table 22: External environment & shocks

	(9)		(10)		(11)		(12)	
	Up	*Down*	*Up*	*Down*	*Up*	*Down*	*Up*	*Down*
Trend growth rate	0.757***	1.368***	0.759***	1.379***	0.761***	1.368***	0.765***	1.374***
	(-7.61)	(6.96)	(-7.74)	(6.68)	(-7.43)	(6.85)	(-7.49)	(6.64)
Years since last shift	0.942***	0.983	0.944***	0.984	0.942***	0.983	0.943***	0.985
	(-3.21)	(-1.41)	(-3.06)	(-1.29)	(-3.29)	(-1.41)	(-3.11)	(-1.27)
Cycle	0.923***	1.173***	0.926**	1.181***	0.921***	1.173***	0.926**	1.178***
	(-2.65)	(4.41)	(-2.47)	(4.24)	(-2.77)	(4.35)	(-2.51)	(4.25)
Growth Volatility	1.031	1.011	1.031	1.012	1.029	1.011	1.028	1.012
	(1.39)	(0.74)	(1.45)	(0.77)	(1.34)	(0.72)	(1.34)	(0.75)
Log(Real GDP p.c. PPP)	0.536***	1.630**	0.509***	1.480**	0.546***	1.645***	0.514***	1.494**
	(-3.60)	(2.52)	(-3.62)	(2.07)	(-3.45)	(2.58)	(-3.48)	(2.09)
Fertility (births per woman)	0.733*	1.416*	0.698**	1.404*	0.740*	1.407*	0.702*	1.390*
	(-1.79)	(1.91)	(-1.96)	(1.69)	(-1.75)	(1.88)	(-1.93)	(1.66)
Pop. ages 0-14 (% total)	0.944*	1.002	0.951	0.995	0.948*	1.003	0.952	0.996
	(-1.81)	(0.05)	(-1.45)	(-0.11)	(-1.71)	(0.06)	(-1.43)	(-0.09)
Pop. ages >=65 (% total)	0.917*	1.061	0.926	1.070	0.920*	1.060	0.923	1.071
	(-1.72)	(0.95)	(-1.44)	(1.04)	(-1.67)	(0.95)	(-1.48)	(1.06)
Polity 2 score (level)	0.978	0.948***	0.981	0.945***	0.979	0.947***	0.980	0.945***
	(-1.07)	(-2.83)	(-0.94)	(-2.71)	(-1.07)	(-2.87)	(-0.98)	(-2.73)
Amelioration Polity 2	0.989	0.780**	0.990	0.784**	0.987	0.778**	0.990	0.782**
	(-0.28)	(-2.06)	(-0.25)	(-2.15)	(-0.32)	(-2.05)	(-0.26)	(-2.12)
Deterioration Polity 2	1.008	1.190**	0.994	1.206**	1.004	1.191**	0.991	1.206**
	(0.18)	(2.39)	(-0.12)	(2.38)	(0.09)	(2.43)	(-0.20)	(2.40)
Duration regime	0.977***	0.979***	0.977***	0.979***	0.978***	0.979***	0.977***	0.979***
	(-3.48)	(-3.73)	(-3.50)	(-3.70)	(-3.44)	(-3.81)	(-3.47)	(-3.79)
Conflict rise >=100%	0.907	2.396***	0.884	2.208***	0.926	2.373***	0.898	2.212***
	(-0.28)	(3.05)	(-0.34)	(2.67)	(-0.22)	(3.02)	(-0.30)	(2.69)
Log(1+inflation)	1.297	1.993*	1.285	2.019*	1.312	2.067*	1.326	2.024*
	(0.68)	(1.77)	(0.67)	(1.79)	(0.72)	(1.89)	(0.76)	(1.81)
Domestic savings (ppt. change)	1.043**	0.947***	1.047**	0.956***	1.043**	0.947***	1.048***	0.957***
	(2.41)	(-3.73)	(2.57)	(-2.85)	(2.37)	(-3.77)	(2.61)	(-2.79)
Domestic credit to private (ppt. change)	1.001	1.034***	1.001	1.036***	1.000	1.033***	0.999	1.035***
	(0.11)	(3.29)	(0.10)	(2.81)	(-0.05)	(3.27)	(-0.09)	(2.77)
Depreciation (LCU/$), win.	1.011**	0.994	1.012**	0.995	1.011**	0.994	1.011**	0.995
	(2.26)	(-0.75)	(2.35)	(-0.71)	(2.25)	(-0.77)	(2.19)	(-0.65)
Depreciation >=100%	0.183**	0.382	0.185**	0.428	0.181**	0.341	0.184**	0.423
	(-2.34)	(-0.89)	(-2.31)	(-0.81)	(-2.31)	(-0.99)	(-2.26)	(-0.79)
Imports gr. - Exports gr.	0.998	1.013**	0.999	1.013**	0.998	1.013**	1.000	1.012**
	(-0.38)	(2.24)	(-0.20)	(2.08)	(-0.29)	(2.21)	(-0.07)	(2.06)
Trade (% of GDP)	1.004**	1.000	1.004**	1.001	1.004**	1.000	1.004**	1.001
	(2.09)	(0.10)	(2.29)	(0.31)	(2.07)	(0.10)	(2.32)	(0.34)
US T-bill (annual change, ppt.)	0.890*	1.089	0.977	1.103	0.873**	1.089	0.928	1.098
	(-1.83)	(1.55)	(-0.30)	(1.59)	(-2.07)	(1.59)	(-0.99)	(1.59)
US T-bill (annual average)	0.918*	1.085***	0.912	1.170*	0.928*	1.092***	0.900*	1.146*
	(-1.86)	(2.60)	(-1.37)	(1.74)	(-1.67)	(2.83)	(-1.70)	(1.77)
Gold price index (annual growth)	0.991	0.997	0.996	1.000				
	(-1.09)	(-0.50)	(-0.43)	(-0.00)				
Gold price index (annual average)	0.976***	1.018***	1.004	1.043***	0.976***	1.020***	1.005	1.039***
	(-3.02)	(3.30)	(0.19)	(3.60)	(-3.09)	(3.57)	(0.29)	(3.78)
Food price index (annual growth)	1.012	1.044***	0.996	1.032**	0.999	1.040***	0.989	1.035***
	(0.77)	(4.10)	(-0.23)	(2.42)	(-0.08)	(4.83)	(-0.82)	(3.46)
Food(growth) x Exporter	1.007	0.968***	1.010	0.969***	1.007	0.968***	1.010	0.969***
	(0.44)	(-2.66)	(0.55)	(-2.65)	(0.44)	(-2.69)	(0.57)	(-2.68)
Food(growth) x Importer	0.993	0.984	0.992	0.984				
	(-0.41)	(-0.82)	(-0.42)	(-0.87)				
Oil price index (annual growth)	0.996	1.000	0.994	1.003				
	(-0.91)	(-0.05)	(-1.38)	(0.70)				
Oil(growth) x Exporter	0.995	0.991**	0.995	0.992*	0.992	0.991**	0.992	0.994
	(-0.69)	(-2.12)	(-0.69)	(-1.86)	(-1.24)	(-2.31)	(-1.32)	(-1.42)
Oil(growth) x Importer	0.995	1.006	0.996	1.006				
	(-0.59)	(1.23)	(-0.48)	(1.22)				
Five Year dummies	No		Yes		No		Yes	
Observations	2464		2464		2464		2464	
Pseudo R^2	0.241		0.259		0.238		0.255	

Exponentiated coefficients; t statistics in parentheses. Standard errors clustered at country level. All variables lagged one year.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Gold is considered as a safe asset and a hedge in turbulent times. Hence, high gold prices are an indication of high risk aversion. More risk averse investors will only lend money at higher rates thereby potentially raising the costs of financing growth-enhancing projects thereby hurting trend growth. Our analysis indicates that although growth in gold prices is insignificant, high averages decrease the relative up-shift risk and increase the relative down-shift risk, with the latter effect being robust to the inclusion of five-year dummies.

Annual averages of gold, food and oil prices are highly correlated: 0.9 correlation between gold and food, 0.94 correlation between gold and oil. In order to avoid multi-collinearity issues, we therefore decided to only focus on food and oil price growth rates and allow the effects to be different for respective exporters and importers.

Rising *Food prices* may increase global food insecurity and poverty, especially in developing countries (cf. e.g. Azzarri et al. [16]), thereby being detrimental to trend growth. Indeed, we find that a one percentage point higher annual growth rate in food prices raises the down-shift relative risk by about 4.4%. This detrimental effect, however, is reversed if the country is a food exporter. In Table 22, the only significant effect of rising *Oil prices* is to slightly protect oil exporting countries from down-shifts.

In further regressions, not shown here but available on request, we investigated several other variables that often appear in the empirical growth literature while keeping the ones that we had already found as significant. The reason for not including these variables in the main specifications presented above is that they reduce the sample size dramatically thereby making results incomparable across specifications, while at the same time often being insignificant.

For instance, we found terms of trade growth to be insignificant while shrinking the sample size from 2464 to 1625. Foreign direct investment net inflows and outflows as a % of GDP are once again insignificant both in levels and ppt. changes, but reduce the sample size to 1768. We investigated the importance of the sectoral composition of the economy by including five variables: the annual growth rates in the value added by the agricultural, manufacturing, and services sectors, and the values added as a % of GDP of the manufacturing and services sectors. The sample size dropped to 1802 and the only effect significant at 10% and robust to the inclusion of five year dummies was that of the growth in the services sector on down-shifts: a one ppt. higher service sector growth increased the relative down-shift risk by about 3.6-3.8%.

Finally, we also looked at short term debt as a percentage of total external debt, finding that a one percentage point higher level increases the down-shift relative risk by almost 5%. The effect is significant at the 0.1% level, and robust to the inclusion of five year dummies. However, unfortunately, including this variable shrinks the sample size from 2464 observations to 1391, mainly because the World Bank provides data on short term debt only for developing countries.

3.4.3 Assessing trend growth prospects

We now wish to turn the insights gained in the previous section about what makes trend growth episodes more or less likely to end next year with an up or a down-shift into a predictive tool that can be used to answer our second motivating question - *"How likely is trend growth to change and in what direction?"* - in real time.

Ideally, we would like to employ all the variables identified as significant in the previous section to give estimates of the conditional probabilities of up and down-shifts in 2016 for most of the countries in our sample. Unfortunately, because of a large number of missing values in 2015, this is not possible. For instance, at the time of writing this paper, the World Bank had updated fertility data for only one country. More annoyingly, some important countries like China and India missed data on either domestic savings or trade.

Although we hope that future research will employ the trend-shifting framework developed in this paper on better datasets and build more interesting and comprehensive predictive models, in what follows, we adopt a less ambitious approach in terms of variables included, but which nevertheless allows us to compare our 2016 hazard estimates for 120 out of the 153 countries for which we have extracted trend growth using iFF.

The up and down-shift hazard estimates reported in Table 25 in the Appendix are constructed using the model shown in Table 23. To get to this model, we re-estimated specification (9) without fertility, domestic savings and the trade variables, then dropped all insignificant effects. Amelioration in Polity 2 score, inflation and growth in oil prices for oil exporters became insignificant, while growth volatility now significantly increases both up and down-shift risks, and a one percentage point higher growth in oil prices decreases the relative up-shift risks around the world by 0.4%. All the other effects go in the same direction as before except for the effect of Pop. ages 0-14 (% total) on down-shifts. A one percentage point larger population in the 0-14 ages bracket now increases the relative down-shift risk by about 7.7%. Note that the protective relationship found in specifications (3) and (4) already became insignificant as we introduced further controls in specifications (5) to (12). However, here the effect

Table 23: Predictive model

	Up	*Down*
Trend growth rate	0.777***	1.271***
	(-8.22)	(6.41)
Years since last shift	0.933***	0.968***
	(-4.22)	(-2.67)
Cycle	0.932**	1.165***
	(-2.56)	(5.87)
Growth Volatility	1.047**	1.025*
	(2.04)	(1.68)
Log(Real GDP p.c. PPP)	0.633***	1.610***
	(-3.51)	(3.67)
Pop. ages 0-14 (% total)	0.913***	1.077***
	(-3.64)	(3.17)
Pop. ages >=65 (% total)	0.897**	1.108***
	(-2.43)	(2.77)
Polity 2 score (level)		0.942***
		(-4.33)
Deterioration Polity 2		1.132*
		(1.75)
Duration Polity regime	0.987***	0.988***
	(-3.07)	(-2.64)
Conflict rise >=100%		2.047***
		(2.65)
Domestic credit to private (ppt. change)		1.038***
		(4.31)
Depreciation (LCU/$), win.	1.010**	
	(2.51)	
Depreciation >=100%	0.349**	
	(-2.08)	
US T-bill (annual average)	0.938*	1.068**
	(-1.79)	(2.47)
US T-bill (annual change)	0.903*	1.114**
	(-1.89)	(2.38)
Gold price (annual average)	0.984***	1.019***
	(-2.97)	(4.05)
Food price (annual growth)		1.035***
		(4.80)
Food price (growth) x Exporter		0.977**
		(-2.21)
Oil price (annual growth)	0.996*	
	(-1.65)	
Observations	3017	
Pseudo R^2	0.197	

Exponentiated coefficients; t statistics in parentheses.
Std. errors clustered at the country level. All variables lagged one year.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 24: Estimated up and down-shift hazards, summary statistics

	Up-shift Hazard	*Down-shift Hazard*
Correlation with Up-shifts	0.188	-0.071
Correlation with Down-shifts	-0.075	0.206
Mean during Up-shifts	0.177	0.043
Median during Up-shifts	0.134	0.022
Mean during Down-shifts	0.035	0.236
Median during Down-shifts	0.018	0.133
Standard deviation	0.107	0.137

goes significantly in the other direction because we are no longer controlling for fertility. Unsurprisingly, the two variables have a correlation of 0.91. Hence, once fertility is excluded Pop. ages 0-14 starts capturing its negative effect discussed above.

Also note that we are not including five-year dummies. The reason is that in this section, unlike the previous one where we were interested in establishing the robustness of our variables to some unobserved time-varying heterogeneity, what we are really looking for is to find a certain combination of *observable* domestic and external variables which, taken together, have historically been significant predictors of up and down-shifts and that we have previously identified as being robust to such unobserved time effects. We could interpret our exercise as wanting to create indices of trend growth instability based on a certain combination of observable domestic and external variables. Unobserved time effects are not something that we can measure in real time and that could help us predict when the current trend growth episodes will end.

The parsimonious predictive model explains almost 20% of trend shifts, which compares well with the more complex specifications we had before and is quite high given the unpredictable nature of such events. It is better at predicting down-shifts than up-shifts since, as shown in Table 24, the estimated up-shift hazards have a 0.188 correlation with the up-shifts, while the correlation between down-shifts and the estimated down-shift hazards is higher: 0.206.[52] This is not surprising since the up-shift hazard is modelled as a function of 14 variables, whereas the down-shift hazard reacts to 17 variables.

The mean up-shift hazard is over four times higher during up-shifts than during

[52]These correlations are based on the whole sample of 5384 observations for which we can compute the hazards and not only the 3017 observations used to estimate the model.

down-shifts (0.177 vs 0.043) and the ratio is around 6 when looking at the median. A more pronounced result holds for down-shifts (6.7 for the mean and 7.4 for the median). However, the mean and median hazard estimates remain quite low, suggesting both that trend shifts are very hard to foretell, and that the model could be improved upon in order to achieve more accurate predictions of trend growth dynamics.

Another interesting way to gauge the performance of the model is to compare the average up and down-shift hazards estimated with the temporal distribution of trend up and down-shifts, analysed in section 3.3. Figure 3.9 superposes the average up and down-shift hazards estimated for each year with the percentage of countries experiencing up and down-shifts in that year. As we can see, the average hazards seem to follow the pattern of the percentages of the respective trend shifts fairly well. In particular, the average down-shift hazard often dominates the average up-shift hazard at the same times as the percentage of down-shifts is larger than the percentage of trend up-shifts and vice versa. More precisely, the correlations (1965 - 2010) between the average hazards and the percentages of countries experiencing shifts are:

	% Countries experiencing	
Average Hazard	*Up -Shifts*	*Down-Shifts*
Up	0.378	-0.579
Down	-0.383	0.629

It is interesting that although the average up-shift hazard is worse at describing the pattern of when up-shifts happen than the down-shift hazard is for the down-shifts: 0.378 correlation vs. 0.629; the up-shift hazard is more sensitive to when down-shifts happen (-0.579 correlation) than is the down-shift hazard for when up-shifts happen (-0.383 correlation).

Note that because we assume a minimum trend growth duration of 5 years when extracting trend growth using iFF, we do not find any trend shifts after 2010. Moreover, all the hazard estimates since 2010 are out-of-sample by construction because none of the observations after 2010 are used in the estimation. The evolution of these out-of-sample hazard estimates suggests that many countries might have experienced down-shifts in this period.

One such country is perhaps Brazil. According to Serrano and Summa [146], Brazil has been living through a challenging period since 2011. Figure 3.10 plots the evolution of the estimated hazards for Brazil, showing that there is indeed a spike in the down-shift hazard in 2011. Table 25 tells us that Brazil has been growing at a trend growth

Figure 3.9 Actual trend shifts & estimated hazards

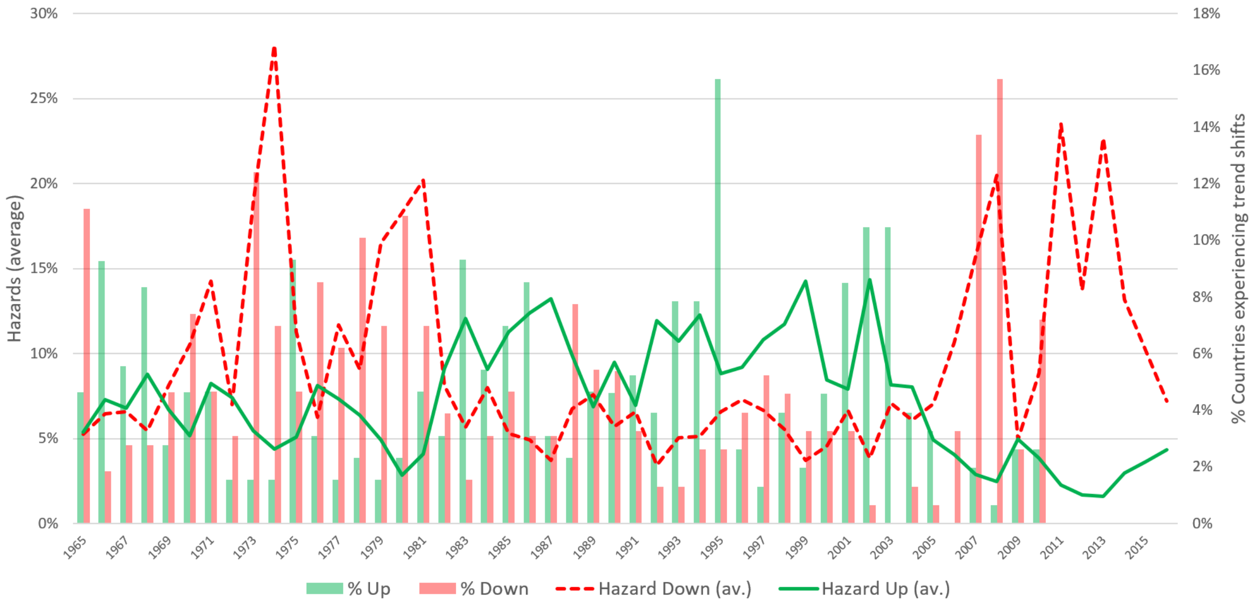
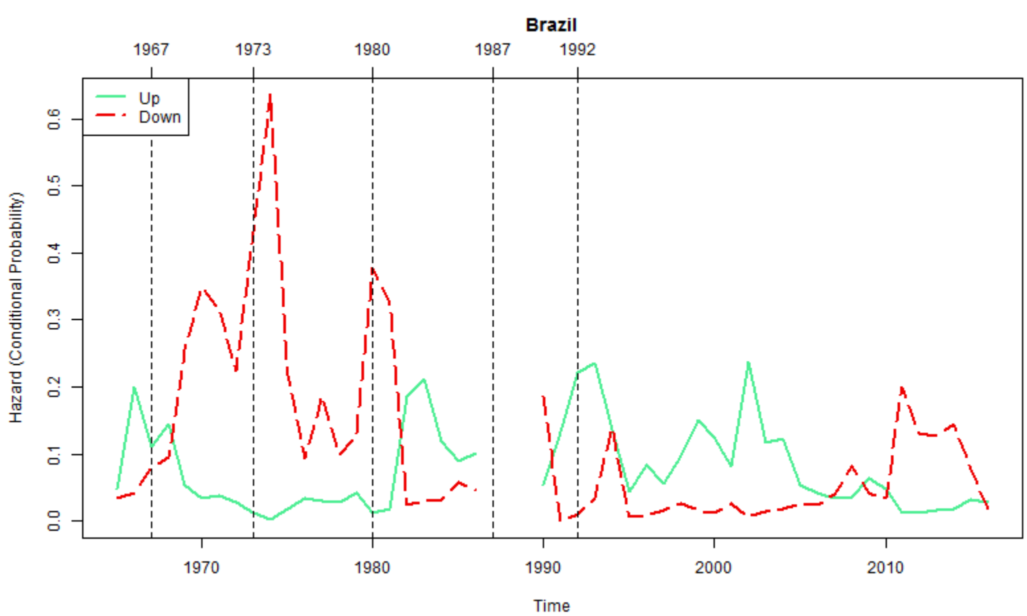


Figure 3.10 Estimated hazards for Brazil



Note: Domestic credit to private sector data missing between 1986 and 1988.

rate of 1.76% p.c. p.a. since 1993 (last trend shift in 1992). Taking this into account, the conditional probabilities of Brazilian trend growth shifting either up or down from this 1.76% trend in 2016 are quite low: 2.8% and 1.6%. A shift in trend growth is a shift in the average growth rate for at least the next five years. These probabilities therefore tell us that it is not very likely that the average Brazilian trend growth over at least the next five years will be very different from 1.76% p.a.

Back in 2011, the rolling estimate of Brazilian trend growth computed using iFF on growth time series up to and including 2010 was 3.29%. As shown in Figure 3.10, in 2011, we would have said that there was a 20% probability that Brazil would grow at a trend growth rate that is at least 3 percentage points lower than this 3.29% trend rate over the five years following 2011, i.e. an average growth rate of close to 0% or below for the years 2012 to at least 2016.[53] Today, in 2016, using data up to and including 2015, we estimate at 1.6% the probability of the event that Brazil experiences a trend down-shift of at least 3 percentage points from its current estimate of trend growth at 1.76%, i.e. grows at a rate of -1.24% p.a. or below between 2017 and at least 2021.

Inspecting the evolution of the hazards visually in Figure 3.10, it seems that down (up) shifts often coincide with spikes in down (up) shift hazards around the shift date, or at least an important rise in the down (up) shift hazard and a fall in the up (down) shift hazard. For instance, for Brazil, 1967 and 1992 are up-shifts, while 1973 and 1980 are down-shifts (cf. Figure 3.8).

This observation is not specific to Brazil. Figure 3.12 shows the trend growth path for Argentina. However, by simply looking at Figure 3.11 we could have already guessed that 1969, 1980, 1997 and 2007 were trend growth down-shifts, whereas 1990 and 2002 were up-shifts.

Argentina's history provides an interesting concrete example which may help us think about how this predictive model could be useful in practice. As shown in Figure 3.11, Argentina's up-shift hazard was rising and the down-shift hazard falling between 2000 and 2003. In 2001, when rating agencies slashed the country's credit rating, the rolling estimate of trend growth was 1.49% p.a. since 1996. The probability of Argentina experiencing a further one percentage point or more down-shift from this trend rate was only 2%, while the probability of the country experiencing an at least three percentage points up-shift for at least the next 5 years was already 8%. In 2002, when Argentina declared default, the trend growth rate estimate (based on data up to

[53]Remember that 3 percentage points is the threshold imposed for a down-shift after an up-shift when extracting trend growth. Figure 3.8 shows that 1992 was an up-shift for Brazil.

Figure 3.11 Estimated hazards for Argentina

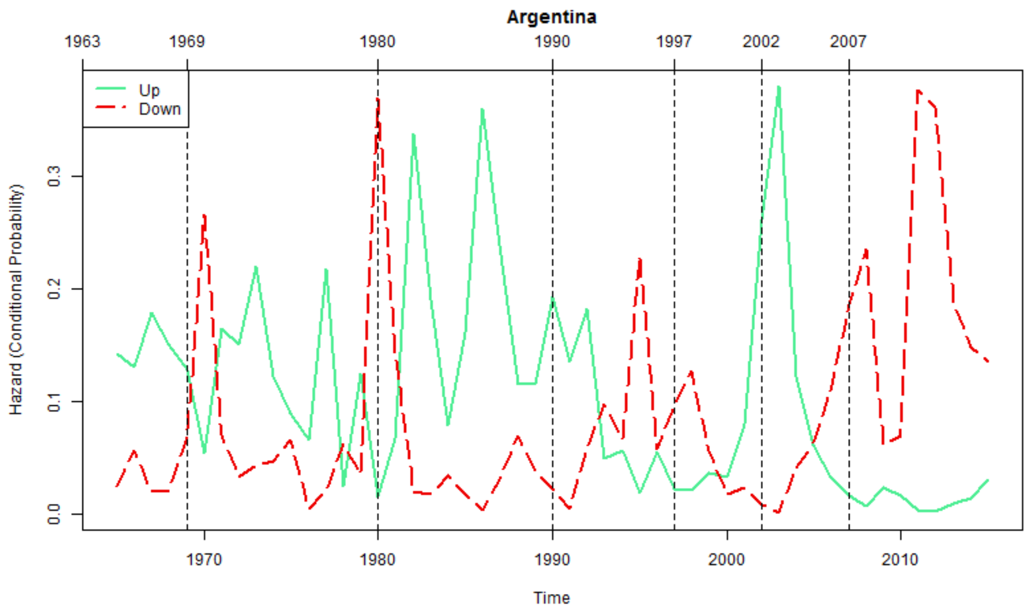
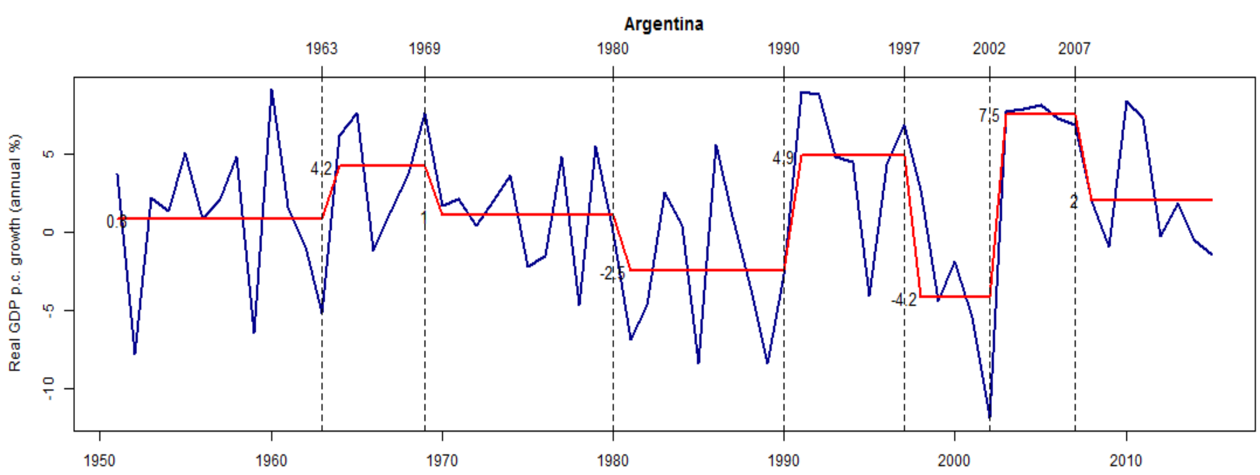


Figure 3.12 Trend growth in Argentina (1951-2015)



Notes: Trend growth (red) extracted from annual real GDP p.c. growth data (blue) using iFF.

and including 2001) slipped to -0.45% p.a. since 1997, while the conditional probability of an up-shift in trend growth rose to 26.2% and that of a down-shift fell to only 0.8%. In 2003, as food prices recovered (annual growth rate of 12.7% in 2002), the up-shift hazard jumped even further to 38%. Hence, as Tenreyro [152] argues: "the 2001-2002 Argentine crisis could and should have been averted [...] had international creditors (and rating agencies) waited for a couple more years", since even a simple model as the one considered here would have pointed to a higher trend growth scenario over the next few years as being much more likely that a further down-shift. And this, even before the food prices started recovering: the 2002 hazards only use 2001 data when food prices actually fell by 3.63%.

As a final example we consider China. Currently, we estimate that China has been growing at 6.81% p.a. since 2008 (last trend down-shift in 2007, cf. Table 25). The conditional probability of a further, at least 1% point, down-shift this year that would make China grow, on average, at 5.8% p.a. or less over at least the next five years is 17.3%. On the other hand, a 3 percentage points trend up-shift is a 1% probability event.

Note that the still very high trend rate at which China is currently growing is certainly contributing to the relatively large down-shift hazard.[54] However, several other factors taken into account in our model are also hampering China's growth prospects. To see this, we can compare China with other countries that are also growing at high trend growth rates currently but have relatively low down-shift hazards. For instance, India has been growing at 5.32% p.a. since 1994 but has a conditional probability of experiencing a trend growth down-shift of only 5.2%. The difference between hazard estimates arises because India has a much higher Polity 2 score than China (9 versus -7). Moreover, in 2015, China experienced a much larger build up of domestic credit to the private sector than India: 13.41% pts. rise in China vs. 0.85% pts. increase in India. Finally, India's growth is also much less volatile China's: standard deviation of 3.24 against 6.15 in China.

In 2016, the average estimated down-shift hazard was still above its up-shift counterpart: 0.07 versus 0.043. However, Figure 3.9 suggests that the down-shift hazard has been on average falling and the up-shift hazard rising since 2013. A relatively mild external environment in 2015 with, for instance, a US 3-month T-bill average rate of only 0.05% that actually fell by 0.03 percentage points throughout the year, has cer-

[54] According to our estimates in Table 25, China is currently ranked fifth in terms of highest trend growth rate, just below Ethiopia (8.02% p.a. since 2004) and slightly overtaking Myanmar (6.80% p.a. since 2011).

tainly aided this trend. However, future rises in US rates will negatively affect growth prospects, while rising commodity prices may negatively impact non-commodity exporters. Moreover, gold prices are still at historically high levels, indicating a high level of risk aversion, which may also be detrimental to trend growth.

3.5 Conclusion & Future Research

Assessing a country's growth prospects is challenging because trend growth is inherently unobserved, and a large number of different domestic and external factors come into interplay at the very same time and could either improve the country's trend growth rate, or worsen it, or counterbalance each other and make the country vibrate around the current trend rate for another several years. Disagreement about a country's growth outlook often arises because different people weight these factors differently and/or use different techniques to extract the trend component from growth time series data.

This paper embraces the basic definition of trend growth as a medium/long term average growth rate and develops a comprehensive empirical methodology which allows us to address the following two questions: *What is the trend growth rate at which a country is currently growing? How likely is it to shift and in what direction?*

The methodology proposed has two components: a trend extraction method that builds on the "Fit and Filter" (FF) approach developed in Kar, Pritchett, Raihan, and Sen (2013) to identify the dates at which trend growth changes significantly, then extract the trend from growth time series as a sequence of medium/long term averages, and an econometric framework, which employs an extended version of discrete-time duration analysis to model and estimate the up and down-shift hazards - the conditional probabilities of up and down-shifts in trend growth next year, conditional on the country having already grown at the current trend growth estimate since the last trend shift.

We employ this methodology on most recent data (up to and including 2015) to give trend growth estimates for 153 countries and up and down-shift hazard estimates for 120 of them. The predictive model developed so far relates the up and down-shift hazards to 20 different variables capturing the current growth environment, the level of development, demographics, institutions, political stability, economic management, and external shocks, and explains almost 20% of trend growth dynamics.

Future research could certainly extend the set of variables considered and build more interesting and comprehensive models.

3.6 Appendix

3.6.1 Estimation

To construct the likelihood function in our case, let's start by considering one specific trend growth episode.

Remember that a trend growth episode starts in the year following the last trend shift $T_{j-1}+1$ or the beginning of the sample if $T_{j-1} = T_0$ - the year in which we have GDP p.c. data for the first time. It ends in the year of the new trend shift T_j or with the end of the sample if $T_j = T$. The likelihood contribution of a trend growth episode depends on how it ends.

Consider, an episode that ends at T_j with a trend up-shift. For this to happen, the country must have survived at a constant trend growth rate from $T_{j-1}+1$ to $T_j - 1$, and conditional on this, experienced a trend up-shift at T_j. Since $\Pr(T_j - T_{j-1} \geq \delta) = 1$, the probability of this scenario reduces to:

$$\begin{aligned}
\mathcal{L}_j^U &= \alpha^U(T_j) \times \prod_{k=T_{j-1}+1+\delta}^{T_j-1} (1-\alpha(k)) \\
&= \frac{\alpha^U(T_j)}{(1-\alpha(T_j))} \times \prod_{k=T_{j-1}+1+\delta}^{T_j} (1-\alpha(k))
\end{aligned}$$

Similarly for the likelihood contribution of a trend growth episode ending with a trend down-shift. In the last case of a trend growth episode ending with the end of the sample, the only thing that we really know is that the trend growth duration of the censored episode is at least $T_j - T_{j-1}$ years with $T_j = T$ this time. Hence its likelihood contribution is simply:

$$\mathcal{L}_j^C = \prod_{k=T_{j-1}+1+\delta}^{T_j} (1-\alpha(k))$$

Putting together $\mathcal{L}^U$, $\mathcal{L}^D$ and $\mathcal{L}^C$, we can write the likelihood contribution of any trend growth episode as:

$$\mathcal{L}_j = \left[\frac{\alpha^U(T_j)}{(1-\alpha(T_j))}\right]^{\mathrm{I}[\Delta\tau(T_j)>0]} \times \left[\frac{\alpha^D(T_j)}{(1-\alpha(T_j))}\right]^{\mathrm{I}[\Delta\tau(T_j)<0]} \times \prod_{k=T_{j-1}+1+\delta}^{T_j} (1-\alpha(k)) \quad (3.18)$$

where I[.] is the indicator function.

Now, let's consider country i with trend shifts $\mathscr{T}_i = \{T_{i1}, ..., T_{im}\}$. The likelihood contribution of its trend growth path is:

$$\mathcal{L}_i = \prod_{T_j \in \mathscr{T}_i} \mathcal{L}_j \tag{3.19}$$

Finally, the likelihood function that we need to maximize is:

$$\mathcal{L} = \prod_{i=1}^{n} \mathcal{L}_i \tag{3.20}$$

where n is the number of countries in our sample.

Taking logarithms and re-arranging:

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{T_j \in \mathscr{T}_i} \sum_{k=T_{j-1}+1+\delta}^{T_j} \{\mathrm{I^U} \log \alpha^U(k) + \mathrm{I^D} \log \alpha^D(k) + (1 - \mathrm{I^U} - \mathrm{I^D}) \log(1 - \alpha^U(k) - \alpha^D(k))\} \tag{3.21}$$

where $\mathrm{I^U} = \mathrm{I}[\Delta\tau(k) > 0]$ and $\mathrm{I^D} = \mathrm{I}[\Delta\tau(k) < 0]$.

The vectors of parameters β^U and β^D can be estimated by substituting the hazard functions 3.13 into eq.3.21 and maximizing it with respect to them.

Several sources in the survival literature, including Allison [11], Jenkins [100], and Allignol et al.[10] explain how such discrete-time competing risk models can be estimated in practice by using standard statistical software for multinomial logits on appropriately re-organized datasets.

Table 25: Trend growth & conditional probabilities of trend shifts

Country	*Trend Growth*	*Last Shift* Year	*Last Shift* Magnitude	*Up-Shift Hazard* Average	*Up-Shift Hazard* 2015	*Up-Shift Hazard* 2016	*Down-Shift Hazard* Average	*Down-Shift Hazard* 2015	*Down-Shift Hazard* 2016
Albania	3.01	2008	-4.26	0.034	0.048	0.049	0.123	0.059	0.049
Algeria	1.94	1994	4.29	0.07	0.02	0.025	0.177	0.072	0.088
Angola	0.75	2008	-7.97	0.036	0.015	0.019	0.272	0.204	0.169
Argentina	2.03	2007	-5.5	0.108	0.031	NA	0.087	0.136	NA
Armenia	0.64	2008	-11.94	0.1	0.1	0.155	0.213	0.09	0.026
Australia	1.92	1951	NA	0.006	0	0	0.012	0.013	0.011
Austria	0.08	2008	-2.06	0.017	0.015	0.017	0.047	0.041	0.034
Azerbaijan	1.05	2009	-19.1	0.161	0.107	0.162	0.332	0.121	0.069
Bahrain	-0.09	1993	-3.89	0.136	0.035	0.044	0.074	0.165	0.067
Bangladesh	4.86	2003	2.11	0.065	0.029	0.029	0.085	0.116	0.081
Belarus	2.2	2008	-8.24	0.058	0.04	0.095	0.304	0.176	0.07
Belgium	1.55	1974	-2.11	0.001	0.001	0.001	0.03	0.053	0.041
Benin	1.31	1960	NA	0.049	0.001	0.001	0.048	0.017	0.011
Bhutan	5.23	1990	-5.87	0.012	0.01	0.01	0.134	0.064	0.059
Bolivia	3.14	2003	1.67	0.1	0.015	0.016	0.065	0.088	0.053
Bosnia and Herzegovina	0.66	2008	-4.03	NA	NA	NA	NA	NA	NA
Botswana	2.83	1989	-7.6	0.019	0.006	0.006	0.063	0.036	0.028
Brazil	1.76	1992	3.62	0.079	0.032	0.028	0.113	0.077	0.016
Bulgaria	0.92	2008	-5.94	0.079	0.031	0.035	0.11	0.069	0.054
Burkina Faso	3.01	1994	1.96	0.052	0.01	0.015	0.083	0.096	0.058
Burundi	4.74	2010	-5.32	0.143	0.013	0.041	0.099	0.113	0.12
Cambodia	4.46	2007	-3.38	0.049	0.03	0.032	0.149	0.246	0.103
Cameroon	2.45	2010	1.34	0.103	0.01	0.031	0.096	0.101	0.105
Canada	1.93	1951	NA	0.009	NA	NA	0.019	NA	NA
Central African Rep.	-6.71	2010	-8.04	0.169	0.42	0.387	0.032	0.01	0.016
Chad	3.31	2010	-1.13	0.117	0.008	0.019	0.143	0.282	0.18
Chile	2.63	1997	-3.51	0.086	0.018	0.017	0.104	0.054	0.046
China	6.81	2007	-3.14	0.039	0.008	0.01	0.173	0.336	0.173
China. Hong Kong	1.96	2007	-4.5	NA	NA	NA	NA	NA	NA
Colombia	3.43	2002	4.19	0.033	0.013	0.019	0.042	0.047	0.038
Comoros	-0.57	1984	-1.55	0.075	0.019	NA	0.028	0.017	NA
Congo	2.13	1999	4.1	0.09	0.009	0.013	0.157	0.199	0.142
Costa Rica	2.63	1983	5.45	0.027	0.003	0.002	0.018	0.015	0.013
Côte d'Ivoire	3.01	2007	4.98	0.102	0.017	0.021	0.078	0.229	0.179
Croatia	-1.5	2008	-5.97	0.074	0.067	0.07	0.102	0.06	0.041
Cyprus	-3.32	2008	-5.86	0.041	0.121	0.106	0.1	0.016	0.041
Czech Republic	0.37	2007	-4.89	0.069	0.029	0.027	0.09	0.16	0.138
D.R. of the Congo	4.69	2009	2.74	0.265	0.016	0.016	0.078	0.202	0.167
Denmark	1.49	1969	-3.36	0.014	0.001	0.001	0.031	0.019	0.015
Djibouti	4.31	2005	3.04	0.106	0.027	NA	0.068	0.087	NA
Dominican Republic	4.32	2004	3.03	0.045	0.01	0.013	0.111	0.185	0.099
Ecuador	1.32	1976	-7.81	0.051	0.005	NA	0.092	0.038	NA
Egypt	0.31	2010	-2.65	0.029	0.051	0.055	0.129	0.204	0.111
El Salvador	1.76	1995	-3.11	0.091	0.015	0.014	0.035	0.034	0.024
Equatorial Guinea	-4.32	2009	-18.91	0.094	0.028	0.208	0.492	0.228	0.012
Estonia	0.56	2007	-8.28	0.057	0.029	0.032	0.19	0.131	0.103
Ethiopia	8.02	2003	8.47	0.15	NA	NA	0.081	NA	NA
Fiji	1.99	1987	3.32	0.069	0.013	0.014	0.063	0.074	0.065
Finland	-1.01	2007	-4.69	0.029	0.015	0.015	0.054	0.044	0.043

Country	*Trend* Growth	*Last Shift* Year	Magnitude	*Up-Shift Hazard* Average	2015	2016	*Down-Shift Hazard* Average	2015	2016
France	-0.06	2007	-1.8	0.017	0.013	0.015	0.083	0.134	0.063
Gabon	3.49	2009	6.43	0.072	0.017	0.022	0.217	0.237	0.177
Gambia	-0.02	1961	NA	0.052	0.002	NA	0.024	0.033	NA
Georgia	4.72	2007	-6.16	0.166	0.026	0.047	0.176	0.208	0.098
Germany	1.25	1991	-1.22	0.029	0.005	0.006	0.071	0.131	0.092
Ghana	4.88	2007	2.79	0.117	0.032	NA	0.068	0.101	NA
Greece	-3.17	2007	-6.67	0.053	0.042	0.037	0.095	0.051	0.06
Guatemala	1.33	1986	4.77	0.049	0.006	0.006	0.07	0.046	0.034
Guinea	-0.88	2008	-1.2	0.058	0.009	0.11	0.036	0.033	0.019
Guinea-Bissau	0.89	2003	3.45	0.13	0.065	0.066	0.051	0.04	0.042
Haiti	2.31	2010	3.67	0.193	0.046	0.096	0.042	0.066	0.079
Honduras	1.49	1986	3.11	0.04	0.011	NA	0.047	0.025	NA
Hungary	0.65	2006	-3.7	0.091	0.023	0.029	0.063	0.16	0.046
India	5.32	1993	3.27	0.038	0.004	0.004	0.033	0.071	0.052
Indonesia	4.09	2001	5.22	0.074	0.019	0.02	0.099	0.139	0.063
Iran (Islamic Rep. of)	0	2007	-5.42	0.142	0.115	NA	0.179	0.232	NA
Iraq	2.05	2008	-10.35	0.017	0.064	0.051	0.475	0.124	0.097
Ireland	-0.15	2007	-3.62	0.01	0.008	0.009	0.044	0.063	0.02
Israel	1.8	1973	-6.24	0.022	0.001	0.001	0.092	0.09	0.068
Italy	-1.29	2007	-1.84	0.026	0.016	0.018	0.06	0.085	0.034
Jamaica	-0.75	2007	-1.07	0.063	0.063	0.054	0.034	0.017	0.014
Japan	0.7	1991	-2.97	0.027	0.003	0.003	0.065	0.055	0.103
Jordan	-0.12	2008	-4.44	0.112	0.046	0.041	0.158	0.104	0.105
Kazakhstan	3.09	2007	-6.32	0.117	0.024	0.031	0.232	0.189	0.124
Kenya	2.64	2003	3.39	0.053	0.019	0.019	0.069	0.149	0.073
Kuwait	-3.43	2006	-9.25	0.15	0.253	0.143	0.244	0.027	0.18
Kyrgyzstan	3.27	1995	16.59	0.155	0.034	0.038	0.055	0.05	0.062
Lao People's DR	5.23	1993	3.37	0.03	NA	NA	0.103	NA	NA
Latvia	0.65	2007	-10.65	0.095	0.032	NA	0.213	0.108	NA
Lebanon	-3.06	2010	-8.95	0.2	0.295	0.241	0.179	0.036	0.06
Lesotho	3.73	2005	1.16	0.032	0.037	NA	0.08	0.055	NA
Liberia	4.98	2007	8.67	0.081	0.038	NA	0.166	0.124	NA
Lithuania	2.59	2007	-7.17	0.073	0.018	NA	0.189	0.151	NA
Madagascar	-1.17	2008	-4.46	0.113	0.123	0.114	0.028	0.039	0.039
Malawi	0.66	2009	-3.1	0.112	0.058	NA	0.094	0.071	NA
Malaysia	2.56	1997	-3.73	0.035	0.016	0.021	0.141	0.111	0.073
Mali	2.01	1984	3.67	0.079	0.006	0.007	0.053	0.081	0.099
Mauritania	1.53	2007	-3.89	0.086	NA	NA	0.157	NA	NA
Mauritius	3.72	1988	-3.13	0.036	0.006	0.007	0.043	0.026	0.033
Mexico	1.15	1986	3.73	0.038	0.01	0.011	0.101	0.059	0.024
Mongolia	8.82	2010	3	0.075	0.013	0.016	0.111	0.13	0.055
Montenegro	0.89	2008	-4.01	NA	NA	NA	NA	NA	NA
Morocco	3.24	1995	3.81	0.04	0.013	0.013	0.088	0.045	0.037
Mozambique	4.46	2001	-4.43	0.043	0.011	0.013	0.13	0.106	0.174
Myanmar	6.8	2010	-4.3	0.249	0.039	NA	0.253	0.417	NA
Namibia	3.34	2003	2.14	0.032	0.014	0.011	0.103	0.09	0.1
Nepal	2.86	1983	2.45	0.052	0.009	0.009	0.039	0.046	0.061
Netherlands	1.52	1973	-2.35	0.016	0.001	0.001	0.032	0.022	0.014
New Zealand	1.22	1966	-3.49	0.016	NA	NA	0.017	NA	NA
Nicaragua	2.51	1993	6.32	0.097	0.017	0.018	0.066	0.08	0.032

Country	*Trend* *Growth*	*Last Shift* Year	Magnitude	*Up-Shift Hazard* Average	2015	2016	*Down-Shift Hazard* Average	2015	2016
Niger	1.87	2004	2.09	0.16	0.023	0.031	0.055	0.087	0.052
Nigeria	3.11	2006	-5	0.134	0.01	NA	0.162	0.305	NA
Norway	-0.33	2007	-3.23	0.008	NA	NA	0.031	NA	NA
Oman	-4.38	2010	-6.84	0.037	0.228	0.151	0.2	0.014	0.049
Pakistan	2.44	1960	2.18	0.029	0.001	0.001	0.057	0.038	0.011
Panama	5.74	2003	4.2	0.043	0.006	0.007	0.139	0.233	0.125
Paraguay	4.99	2009	1.85	0.062	0.018	0.025	0.1	0.119	0.09
Peru	4.37	2002	3.8	0.098	0.018	0.018	0.087	0.142	0.135
Philippines	3.76	2002	2.7	0.068	0.015	0.013	0.06	0.075	0.102
Poland	4.18	1993	6.87	0.073	0.007	0.008	0.071	0.069	0.053
Portugal	0.01	2001	-2.23	0.041	0.014	0.016	0.093	0.039	0.032
Qatar	1.6	2001	-5.93	0.168	0.042	0.041	0.109	0.154	0.088
Republic of Korea	3.63	1996	-4.74	0.034	0.01	0.011	0.19	0.132	0.091
Republic of Moldova	2.99	2008	-3.68	0.336	0.102	0.187	0.036	0.039	0.026
Romania	1.25	2008	-6.14	0.074	0.033	0.036	0.12	0.096	0.072
Russian Federation	0.33	2008	-6.89	0.144	0.069	0.146	0.186	0.091	0.026
Rwanda	4.29	2008	-1.29	0.069	NA	NA	0.113	NA	NA
Saudi Arabia	2.22	2002	4.16	0.077	0.009	0.008	0.159	0.114	0.182
Senegal	1.32	1993	3.93	0.084	0.013	0.015	0.045	0.046	0.039
Serbia	0.17	2008	-6.03	0.052	0.088	0.091	0.128	0.053	0.084
Sierra Leone	3.41	2004	-4.8	0.188	0.014	0.113	0.074	0.145	0.003
Singapore	2.68	1997	-3.8	0.035	0.009	0.009	0.179	0.096	0.062
Slovakia	1.45	2008	-5.52	0.068	0.043	NA	0.098	0.072	NA
Slovenia	-0.91	2008	-5	0.053	0.038	0.044	0.087	0.108	0.046
South Africa	1.51	1993	3.18	0.032	0.017	0.019	0.039	0.029	0.017
Spain	-0.75	2007	-2.82	0.037	0.029	0.028	0.084	0.04	0.04
Sri Lanka	6.05	2004	2.07	0.04	0.012	0.01	0.063	0.291	0.324
Sudan (Former)	-1.22	2010	-5.02	0.094	0.073	0.082	0.123	0.105	0.066
Swaziland	1.26	1990	-4.44	0.019	0.008	0.009	0.108	0.094	0.067
Sweden	2.15	1951	NA	0.01	0	0	0.034	0.011	0.015
Switzerland	0.86	1973	-2.26	0.005	0	0	0.008	0.007	0.005
Syria	-8.36	2009	-9.98	0.086	NA	NA	0.169	NA	NA
Taiwan	3.87	1995	-3.08	NA	NA	NA	NA	NA	NA
Tajikistan	3.07	2008	-3.42	0.061	0.048	0.084	0.174	0.144	0.053
TFYR of Macedonia	1.99	2008	-3.09	0.119	0.047	0.056	0.061	0.136	0.049
Thailand	3.67	2001	4.4	0.067	0.037	0.028	0.157	0.204	0.218
Togo	2.54	2009	3.48	0.124	0.035	0.042	0.062	0.104	0.08
Trinidad and Tobago	0.57	2006	-8.91	0.065	0.028	0.028	0.073	0.038	0.026
Tunisia	0.82	2010	-1.86	0.041	0.095	0.107	0.131	0.098	0.041
Turkey	1.92	2006	-3.85	0.053	0.044	0.041	0.105	0.132	0.198
Turkmenistan	9.56	2004	4.78	NA	NA	NA	NA	NA	NA
Tanzania	3.37	1998	2.38	0.051	0.008	0.011	0.083	0.145	0.13

Country	*Trend Growth*	*Last Shift* Year	Magnitude	*Up-Shift Hazard* Average	2015	2016	*Down-Shift Hazard* Average	2015	2016
Uganda	1.48	2010	-3.5	0.101	0.022	0.034	0.112	0.274	0.139
Ukraine	-1.58	2007	-9.95	0.249	0.242	0.317	0.113	0.036	0.013
United Arab Emirates	2.43	2010	11.47	0.211	0.093	0.101	0.088	0.087	0.05
United Kingdom	2	1951	NA	0.005	0.004	0	0.02	0.019	0.004
United States	2.03	1951	NA	0.001	0	0	0.008	0.002	0.001
Uruguay	4.91	2003	8.04	0.09	0.008	0.009	0.092	0.141	0.108
Uzbekistan	6.57	2003	3.91	NA	NA	NA	NA	NA	NA
Venezuela	-2.1	2008	-10.77	0.08	NA	NA	0.084	NA	NA
Viet Nam	5.65	1991	2.28	0.018	0.006	0.006	0.092	0.065	0.062
Yemen	-4.57	2010	-6.5	0.046	NA	NA	0.138	NA	NA
Zambia	4.38	2002	3.22	0.094	0.009	0.011	0.092	0.124	0.1
Zimbabwe	12.13	2008	17.54	0.135	NA	NA	0.069	NA	NA

Notes: Trend growth estimates based on data up to and including 2015, except for Syria (growth data stops in 2014). Hazard estimates based on model in Table 23. Trend rates extracted using iFF with the following assumptions:
Minimum trend growth duration: 5 years
Trend shifts significance filter: 1% pt. change for a shift in the same direction. 3% pts. change for a shift in the opposite direction.
Example of how to read the table: China has been growing at 6.81% p.a. since 2008 (last trend shift in 2007). 2007 was a trend down-shift, hence a further down-shift would be identified as a $\leq$ 5.81% p.a. average growth rate over the next $\geq$ 5 years, whereas an up-shift would be a $\geq$ 9.81% p.a. average growth rate over the same period. In 2016, the conditional probabilities of trend down and up-shifts are 17.3% and 1% respectively.

4 Credit Risk Spillovers, Systemic Importance and Vulnerability in Financial Networks

Abstract

How does the change in the creditworthiness of a financial institution or sovereign impact its creditors' solvency? We address this question in the context of the Eurozone sovereign debt crisis. Considering the network of Eurozone member states, interlinked through investment cross-holdings, we model default as a multi-stage disease with each credit-rating corresponding to a new infection phase, then derive systemic importance and vulnerability indicators in the presence of financial contagion, triggered by the change in the creditworthiness of a network member. We further extend the model to analyse not only negative, but also positive credit risk spillovers.

4.1 Introduction

Why does the downgrade of a European country not only raise the CDS spreads of that specific sovereign but also those of other Eurozone member states (Arezki et al. [13])? More generally, how does the change in the creditworthiness of a financial institution or state impact its creditors' solvency?

This paper addresses such questions, investigating how a financial event that originates in one specific country can spread beyond its borders, infecting other states like an epidemic. We model default as a process - a multi-stage disease. Each credit-rating corresponds to an infection phase, during which default happens and therefore contagion is transmitted with a certain probability (Section 4.2). The model is general and could be applied not only to sovereigns, but to any financial institutions such as banks, firms, etc. interlinked by mutual financial liabilities. However, as an example, we illustrate its workings in the context of the European sovereign debt crisis. The seventeen member states are the nodes in the network, and the weighted directed edges between them measure cross-country investment flows (Fig. 4.1).

Our goal is to develop indicators of node importance and vulnerability by investigating how the exogenous change in the creditworthiness of one of the financial network members impacts the creditworthiness of all the other ones (Section 4.3). Firstly, we derive some analytical indices from the early time properties of the model, then employ computer simulations to measure *systemic* importance and vulnerability. These latter indicators are systemic in the sense that they not only capture the immediate effect of the "exogenous risk that hits the system" (such as an aggregate exogenous shock or an idiosyncratic shock to one of the nodes in the network), but also the "endogenous risk generated by the system itself" (Zigrand [166]). The system here is the network and the aim is to find indicators that will take into account how the structure of the links shapes and propagates an initial exogenous shock. We also extend the model to analyse how the same feedback mechanisms that generate endogenous risk, could set off a process of positive contagion, reversing negative market sentiments (Section 4.4).

This paper can be related to the existing vast literature on financial contagion. At the two extremes, contagion is classified as *pure* - when a herd of investors drives apparently healthy and unrelated economies towards sunspot equilibria - or *fundamentals-based* (Masson [121]). In reality, however, spillovers are complex and encompass both features: they spread through real or financial channels, while still retaining some randomness driven by market sentiments. We try to capture this duality by, on the one hand, considering a financial network of cross-country investment flows, while on the other hand making downgrades happen stochastically. Indeed, the probability that an agent transmits the "default disease" to its network neighbours depends on both: its actual credit rating (an agent with a lower credit-rating, i.e. in a more advanced infection phase, transmits the negative contagion at a higher rate) and the interaction intensities (a stronger mutual relation increases the transmission probability).[55]

One of the first models studying financial contagion on networks is Allen & Gale [9]. The authors extend Diamond & Dybvig [46] to a four banks system, showing how the completeness and distribution of interconnections determine the extent of spillovers following a bank-specific shock. With evenly allocated deposits, contagion may be completely avoided, whereas in an incomplete system, a cascade of failures might emerge. Allen & Babus [8] give an overview of some recent developments in this field. Espinosa-Vega & Sole [55] build an interbank exposure model, simulating credit and liquidity shocks. Their algorithm starts with the default of a country's banking system shifting the balance sheets of yet solvent banks and triggering new failures. In a similar spirit, Elliott et al. [53] construct a theoretical model in which the market values of organisations are interdependent through the network of cross-holdings. The default of an organisation (bank or country) changes the values of all the other ones inducing those, whose new values fall below certain specified "bankruptcy thresholds", to fail as

[55] For a discussion of the relative importance of trade linkages versus macroeconomic similarities in currency crises, see Eichengreen et al. [50]. Gerlach & Smets [80], Corsetti et al. [41], and Pesenti & Tille [130] present theoretical models of contagious transmission with applications to the Asian currency crisis and the ERM turmoil.

well. The contagion process continues until either the algorithm converges with no new failures, or no solvent organisation remains.

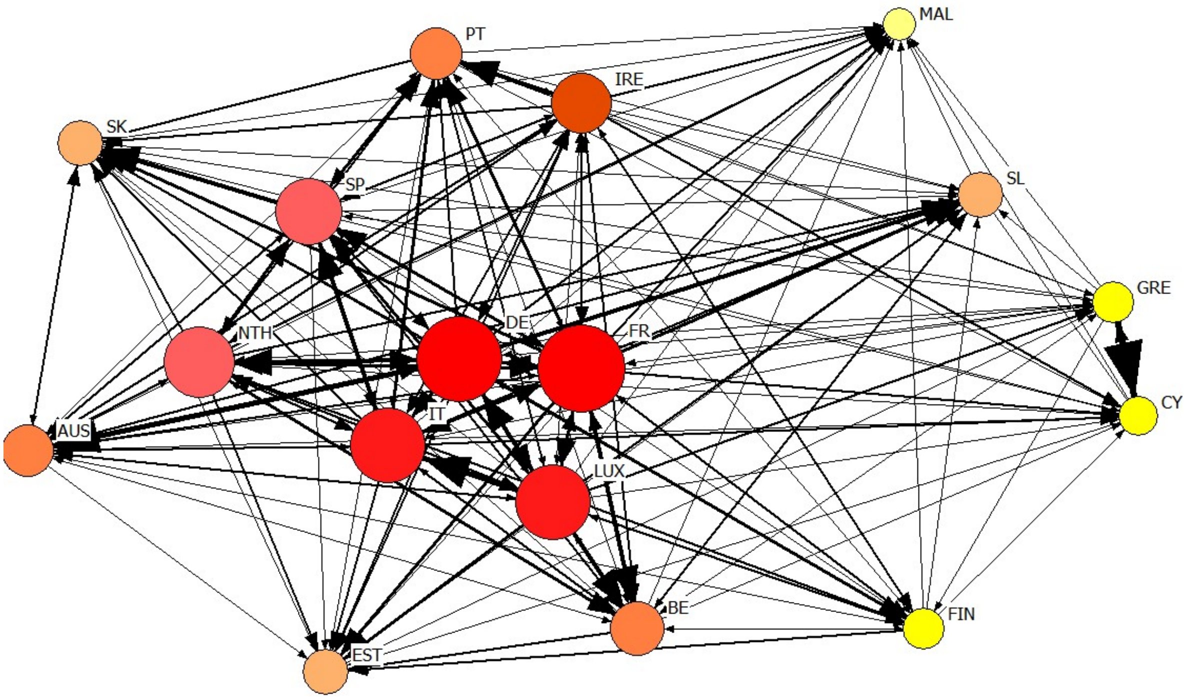
One of the main criticisms of such studies has been their limited scope given the "extremely rare" nature of "contagious *failures*" (Upper [156]). Here, instead of analysing how financial contagion results from the *initial default* of a bank or country, we model the *default process* itself, and investigate how financial contagion can be triggered by simple *changes in the creditworthiness* of one of the network members.

The credit-rating determination of a bank (Bissoondoyal-Bheenick & Treepongkaruna [27]) or a sovereign (Melliosa & Paget-Blanc [124], Afonso [3], Cheung [37]) have been traditionally studied using econometric ordered-response models with creditworthiness as the latent variable. The problem with this approach is that it completely ignores the possibility of credit risk spillovers between financial entities. Unfortunately, once we model agents' interactions explicitly through a network, and admit that their creditworthiness levels are interrelated and determined simultaneously, this econometric framework can no longer be used because of endogeneity.

We propose a different approach to studying creditworthiness and credit risk spillovers that takes inspiration from the epidemiology literature. Indeed, the highly interconnected and complex nature of the global financial system has spurred researchers to draw interesting and original parallels between financial networks, ecosystems (May et al. [122]), epidemiology, and even engineering (Haldane [86]). Demiris et al. [45] explore financial contagion in the context of a Susceptible-Infected-Recovered (SIR) model, emphasising its advantages over more conventional modelling approaches. In particular, the SIR framework allows them to model explicitly the country-interdependencies that are essential to the propagation of a crisis, measure crisis severity by a threshold parameter instead of composite macroeconomic indicators as in Kaminsky & Reinhart [106], and evaluate potential policy interventions.

Beside these theoretical advances, an increased interest in complexity economics and agent-based modelling have recently re-emphasized the usefulness of computer simulations in understanding complex systems (Farmer & Foley [56]). For instance, Caporale et al. [32] develop a multinomial model using time series data on stock returns during the East Asian crisis (1997), and thereby disentangle potentially destabilizing connections that could signal the inception of a contagion process. Gai et al. [77] experiment with different parameter configurations, studying how complexity and concentration affect the resilience of a financial system. We follow this trend and use computer simulations not only to derive indicators of systemic importance and vulnerability, but also

Figure 4.1 Eurozone network, 2011 Total Portfolio Investment (TPI) shares



Notes: The weighted and directed link from country j to country i (v_{ij}) is the proportion of country i's total investment flowing into j. Nodes coloured and weighted by their average degrees $\frac{1}{2}\left\{\sum_i v_{ij} + \sum_j v_{ij}\right\}$.

to incorporate positive contagion, thereby making the model more complex, interesting, and realistic.

4.2 Credit Risk Spillovers

4.2.1 The determinants of creditworthiness

Consider a network with $n \in \mathbb{N}$ nodes. As an illustration, Figure 4.1 depicts the network of the seventeen Eurozone member states. Since the ultimate goal is to study credit risk spillovers, the links between nodes should reflect the intensity of potential contagion flows. Fig. 4.1 looks at the 2011 cross-border *Total Portfolio Investment* (TPI) flows. The data is available from the IMF (Consolidated Portfolio Investment Survey). TPI flows are reported on an annual basis and include long & short term equities and debt-securities. We transform the data into investment shares since absolute values of investment flows vary with the overall size of the economy and need to be normalised. Formally, we define the weighted, directed link from j to i (v_{ij})

as the proportion of country i's total investment flowing into j. Intuitively, the arcs follow contagion flows so that the larger v_{ij}, the more significant the direct potential spillover from j's downgrade onto i's creditworthiness. In Fig. 4.1, nodes are coloured and weighted by their average degrees $\frac{1}{2}\left\{\sum_i v_{ij} + \sum_j v_{ij}\right\}$.

Investment shares v_{ij} can be summarized in an $n \times n$ adjacency matrix $\mathbf{V}$ (Table 26). Note that $v_{ii} = 0$ (foreign investment only) and row i gives the distribution of country i's TPI across other Eurozone member states. For instance, Austria invests 21.46% into Germany and only 0.64% into Greece. Rows do not sum up to 100% because countries also invest in non-EZ states. In general, $v_{ij} \neq v_{ji}$, e.g. Italy invests 26.22% into Luxembourg, while Luxembourg only invests 4.64% into Italy.

Let the creditworthiness of each node at time t - y^*_{it}- be described by:

$$y^*_{it} = \mathbf{x}^{'}_{\mathbf{it}}\boldsymbol{\beta} + \omega \sum_j v_{ij} y^*_{jt} \tag{4.1}$$

where $\boldsymbol{\beta}$ is a vector of parameters, and $\mathbf{x_{it}}$ - a vector of economic indicators that include variables like debt ratios, growth and inflation rates, default history, etc. in the case of sovereigns. If banks were considered instead, $\mathbf{x_{it}}$ would include variables like asset quality, liquidity risk, capital adequacy, operating performance, etc. The parameter ω captures how changes in i's debtors' creditworthiness will affect its own creditworthiness given i's portfolio allocation $\sum_j v_{ij}$. The model could be made more general by allowing all parameters to depend on i, reflecting the fact that different characteristics do not necessarily have the same importance in determining the creditworthiness of two different financial entities. However, this would make the model more complex, and we leave this extension to future research.

Even though in practice many credit-rating agencies have been recently accused of assigning credit-ratings that do not reflect the true creditworthiness of a bank or sovereign, theoretically we can assume that there exists a direct mapping from creditworthiness to credit-ratings. Credit-ratings are ordinal qualitative variables often designated by alphabetical letters. Standard & Poor's ratings for example range from AAA (no default risk) to C - the worst possible rating before the restricted default D. Suppose there are k possible ratings in total (the actual number varies from agency to agency). We can translate credit-ratings to a numerical scale with the highest possible rating denoted as k and the lowest possible one as 0. Financial institution i is downgraded as its creditworthiness y^*_{it}, defined by eq.4.1, drops below certain thresholds α:

Table 26: Adjacency matrix, 2011 Total Portfolio Investment (TPI) shares (%)

	AUS	BE	CY	EST	FIN	FR	DE	GRE	IRE	IT	LUX	MAL	NTH	PT	SK	SL	SP
AUS	0	1.319	0.140	0.023	1.368	8.524	21.456	0.638	3.705	7.398	8.434	0.010	7.327	0.525	0.790	0.710	3.538
BE	2.265	0	0.032	0.002	0.838	18.459	6.689	0.455	5.700	4.845	19.111	0.006	14.661	0.730	0.195	0.226	5.172
CY	0.496	1.838	0	0	0.411	4.291	2.372	27.833	3.708	3.448	2.419	0	2.121	0.300	0.003	0.112	0.924
EST	3.113	3.352	0.311	0	7.328	14.679	7.328	0.096	6.346	7.807	11.997	0.024	8.357	0.431	0.455	0.216	1.461
FIN	1.425	0.668	0.036	0.094	0	8.007	10.144	0.291	5.159	2.243	8.755	0	5.742	0.398	0.022	0.079	2.156
FR	3.072	4.340	0.012	0.001	0.678	0	11.303	0.587	4.138	10.761	6.104	0.038	11.883	1.381	0.039	0.104	8.281
DE	3.865	1.415	0.042	0.002	1.053	12.668	0	0.516	5.263	7.299	16.218	0.004	9.968	0.935	0.180	0.127	6.878
GRE	0.929	0.330	3.393	0	0.052	3.284	2.727	0	2.394	1.857	7.902	0	2.242	0.296	0	0	0.641
IRE	0.358	0.544	0.009	0.008	0.594	5.374	7.174	0.144	0	5.724	2.248	0.011	3.820	2.633	0.017	0.009	2.031
IT	1.719	0.810	0.007	0	0.328	13.094	10.381	0.430	8.634	0	26.215	0.022	6.483	0.690	0.035	0.116	4.505
LUX	1.036	1.792	0.045	0.009	0.648	9.750	12.381	0.103	2.711	4.635	0	0.019	5.164	0.264	0.019	0.023	2.299
MAL	1.939	1.011	0.107	0	0.308	3.942	5.765	0.950	5.191	2.309	2.720	0	4.851	1.134	0.175	0.942	2.677
NTH	2.255	1.673	0.015	0.002	1.181	11.271	15.964	0.228	3.445	3.487	5.355	0.004	0	0.807	0.037	0.024	3.920
PT	1.372	1.762	0	0	0.416	9.599	5.298	1.327	20.645	9.603	6.566	0	7.113	0	0.050	0.028	11.795
SK	5.286	1.173	0.527	0	2.431	9.394	4.222	1.154	8.663	7.995	2.864	0	6.088	4.343	0	2.444	15.855
SL	6.269	3.923	0.040	0.015	0.805	13.337	18.605	2.721	2.348	8.654	4.082	0	7.023	0.953	1.227	0	2.259
SP	1.678	2.054	0	0	0.358	13.227	8.685	0.872	6.806	15.753	9.975	0	9.567	3.933	0	0	0

Source: Coordinated Portfolio Investment Survey (CPIS), IMF

cf. Table 8: Geographic Breakdown of Total Portfolio Investment Assets: Total Portfolio Investment, http://cpis.imf.org/

The adjacency matrix is obtained by dividing the original entries in the CPIS matrix by the total value of investment for each country.

$$cr_{it} = \begin{cases} k & \text{if } y^*_{it} > \alpha_k \\ k-1 & \text{if } \alpha_{k-1} < y^*_{it} \leq \alpha_k \\ \vdots & \\ 0 & \text{if } y^*_{it} \leq \alpha_1 \end{cases}$$

where cr_{it} is i's credit-rating at time t.

As mentioned in the introduction, eq.4.1 cannot be estimated econometrically through an ordered-response model because the sum $\sum_j v_{ij} y^*_{jt}$ - which in the context of the illustrative Eurozone network can be interpreted as country i's creditworthiness-adjusted portfolio investment - is unobservable. Since, in practice, credit-ratings are only an imperfect measure of the underlying creditworthiness, replacing y^*_{jt} by j's observed credit-rating at time t in eq.4.1 could make things even worse because of the second important problem: endogeneity. Indeed, the whole idea of this paper is to argue that y^*_{it} and y^*_{jt} will be determined simultaneously if v_{ij} and v_{ji} are non-zero.

To identify credit risk spillovers, we therefore take inspiration from the epidemiology literature and model the default process as a multi-stage disease with each credit-rating corresponding to a new infection phase.

4.2.2 The default process

From the "Susceptible - Infected" (SI) model to the "Solvent-Default" (SD) one The SI model was initially developed to analyse human disease spreading on contact networks (Kermack & McKendrick [110]). Although the process of infection is much more complex than this two-states model, it remains a "useful simplification" to the extent that it captures the contagion dynamics "happening at the level of networks" (Newman [128]).

At any point in time, an individual is either susceptible or infected. Suppose that you are susceptible - which happens with probability s_i. To catch the disease, one of your neighbours must already be infected - the probability of this event is d_j. Since this infected neighbour transmits the disease at rate δ, the probability that you become infected at any point in time is:

$$\frac{dd_i}{dt} = \delta \sum_j v_{ij} \{s_i d_j\} \tag{4.2}$$

where the accolades on the right-hand side take into account the correlations (joint probabilities) for nodes i and j to have the specified states, e.g. $\{s_i d_j\}$ is the "average probability that i is susceptible and j is infected at the same time" (Newman [128]). Interaction intensities v_{ij} play a crucial role in eq. 4.2. Two individuals, interacting with exactly the same people, could have completely different infection probabilities if the first person's contacts mainly include already contaminated individuals, whereas the second one is more interlinked with still healthy people.

Translating the SI framework into a "Solvent-Default" (SD) one is rather trivial. We can simply think of the interaction intensity v_{ij} as the strength of the potential contagion flow from j to i. For instance, in the Eurozone network illustrated in Fig. 4.1, the v_{ij}'s will be the TPI shares. Eq. 4.2 implies that two countries, Z and W, investing in exactly the same sovereigns with the only difference that $\sum_j v_{Zj}$ puts relatively higher weights on countries with larger d_j than $\sum_j v_{Wj}$, would have very different default probabilities. In particular, Z's default probability would be larger than W's. Notice that this discrepancy would occur not because of a difference in d_j, but as a consequence of different TPI shares' distributions.

Unfortunately, this simple SD framework is inappropriate for the investigation of *credit risk spillovers*, because it only models two states: Solvent and Default. In reality, a solvent financial institution does not suddenly declare bankruptcy; it undergoes a process of rating downgrades, which we shall model as a multi-stage disease.

Default as a multi-stage disease Modeling default as a multi-stage disease makes the previous framework more complex, but allows financial institutions to differ in *how* solvent they are.

We can think about each credit-rating downgrade as marking the beginning of a new infection phase. Let $\theta(cr_{it}) \in [0; 1]$ be the rate at which institution i with credit-rating cr_{it} transmits the "default disease" to its creditors. An institution with lower creditworthiness, i.e. in a more advanced infection phase, is more likely to default itself. It should therefore be more virulent and transmit the default contagion at a higher rate, i.e. $\theta(cr_{it})$ must be decreasing in cr_{it}.

For simplicity, suppose there are only four possible credit-ratings ($k = 4$): A, B, C, and D. Let a_{it}, b_{it}, c_{it} and d_{it} denote the probabilities that institution i has rating A, B, C, or D respectively at time t. Clearly:

$$a_{it} + b_{it} + c_{it} + d_{it} = 1$$

because an institution must be rated at any point in time.

Further, let the transmission rates be: $\theta(A) = 0$ and $\theta(cr) \in (0; 1]$ for any $cr \neq A$.

Given the network of n financial institutions or states, i can be downgraded from A to B for two reasons. Firstly, as a result of a credit risk spillover from one of its debtors. For this, i must start with rating A which happens with probability a_{it}. Moreover, one of its debtors must have rating $cr_{jt} \in \{B,\ C,\ D\}$ - which happens with probabilities $z_{jt} = \{b_{jt},\ c_{jt},\ d_{jt}\}$ respectively - and transmit the contagion at rate $\theta(cr_{jt})$. Secondly, i could also be downgraded because of an exogenous deterioration in one of its characteristics x_{it}. The following differential equation describes the probability that i loses credit-rating A:

$$\frac{da_i}{dt} = -\omega \sum_j v_{ij} \{a_{it} z_{jt}\} \, \theta(cr_{jt}) + \beta^A \frac{dx_i}{dt} \tag{4.3}$$

Similarly, the probability that i has rating B increases in i's probability of being downgraded from A to B, but decreases in i's probability of being downgraded from B to C:

$$\frac{db_i}{dt} = \omega \sum_j v_{ij} \{a_{it} z_{jt}\} \, \theta(cr_{jt}) - \omega \sum_j v_{ij} \{b_{it} z_{jt}\} \, \theta(cr_{jt}) + \beta^B \frac{dx_i}{dt} \tag{4.4}$$

The differential equation for rating C is described in a similar way:

$$\frac{dc_i}{dt} = \omega \sum_j v_{ij} \{b_{it} z_{jt}\} \, \theta(cr_{jt}) - \omega \sum_j v_{ij} \{c_{it} z_{jt}\} \, \theta(cr_{jt}) + \beta^C \frac{dx_i}{dt} \tag{4.5}$$

Finally, the last downgrade from C to D corresponds to the default after which the country exits the system:

$$\frac{dd_i}{dt} = \omega \sum_j v_{ij} \{c_{it} z_{jt}\} \, \theta(cr_{jt}) + \beta^D \frac{dx_i}{dt} \tag{4.6}$$

where the parameter on the exogenous shock β is allowed to depend on the credit rating under consideration.

There is one important difference between this extended SD model and the existing multi-stage disease ones from the epidemiology literature. In the latter, "the only stochastic step [...] is transmission of the disease" (Jaquet & Pechal [99]) - i.e. once an individual becomes infected, she will traverse all the disease stages deterministically until reaching the final phase (e.g. becoming resistant) - whereas in the multi-stage SD model, progression towards the next infection phase still depends on contact with already contaminated units.[56][57]

4.2.3 Computer simulations

The model presented in the previous section has no closed form solution. To analyse it, we have written a graph-based computer algorithm whose main component is a function called *spillovers*. This function takes as arguments an adjacency matrix $\mathbf{V}$, a transmission-rates vector $\boldsymbol{\theta}$, and the vector of initial ratings rat_0. Firstly, it associates to each country i a transmission rate $\theta(cr_i)$ that depends on its rating. After computing country-specific downgrade probabilities using eq.4.7, some countries are stochastically downgraded:

$$\mathbf{Pr}(i\ downgraded) = \omega \sum_j v_{ij}\theta(cr_j) \tag{4.7}$$

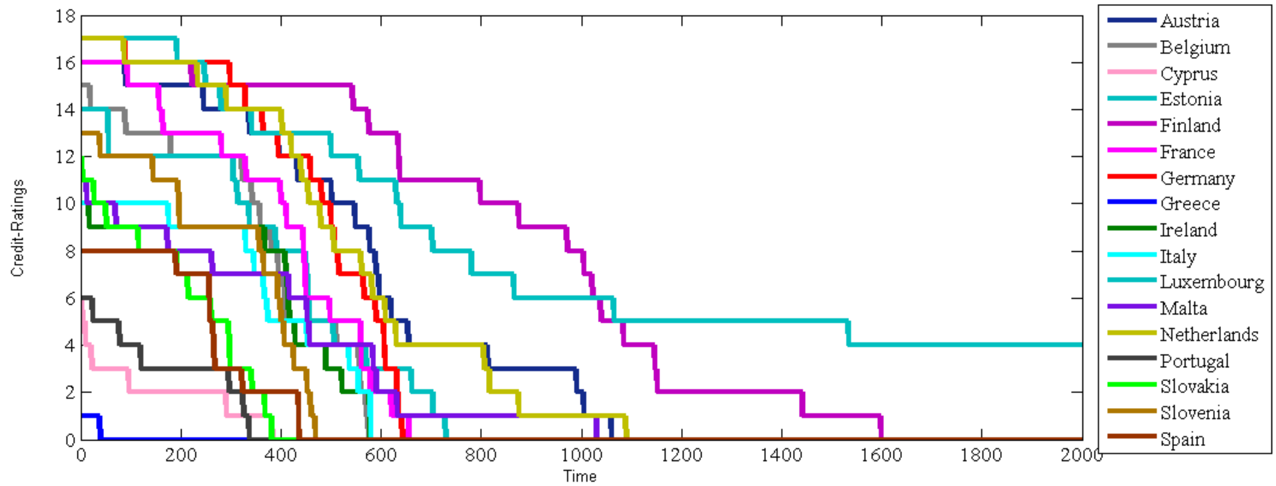
Here, $\omega \in \mathbb{R}^+$ captures the overall spillovers' magnitude (the "severity" of the disease) and allows to change the speed of the simulations without altering downgrade rankings.

Spillovers' output is a new vector of ratings: rat_1. The *contagion* function recalls *spillovers* until all countries default - which happens at time T_d - using the *spillovers'* previous ratings output vector as the new input at each iteration step, and producing a matrix with columns: rat_0, rat_1, $rat_2, ...,$ rat_{T_d}. Technically, one country usually remains solvent. Since the simulation stops when all but one country default, this solvent country also defaults at T_d.

[56]Note that the infection phases of the node's debtors do not have to be "more advanced" than its own contamination phase in order to make him progress to the next infection stage. Consider the following example: country X currently has rating B and has only two debtors: countries Y and Z, both with ratings A, i.e. in "less advanced" infection phases than itself. If country Y is downgraded from A to B, this will increase the probability of country X being downgraded to C, i.e. there will still be a credit-risk spillover despite the fact that country Y was initially less infected than country X.

[57]Aside from the economic content of this paper, there is a need to properly understand the dynamics and phase diagrams of a multi-stage epidemic model in which progression to the next infection phase depends on interaction with other contaminated people. Although such a model could not be applied in common epidemic settings such as the transmission of a viral infection, it could be very useful in understanding contagious mental illnesses. For instance, Joiner & Katz [104] investigate 40 studies conducted between 1976-1997 "that examined the relationship between two non-genetically related individuals' levels of depression or negative mood", and find evidence of significant contagion for all the 12 studies that analysed syndromal *depression spillovers* between "college roommates, dating couples, young spouses, elderly spouses, and relatives." Interestingly, they also find that only 13 of the 28 studies dealing with *negative mood*, report significant spillovers, concluding that "depressive symptoms are [...] more contagious than negative mood." This agrees with the framework used in this paper, whereby more advanced infection phases are presented as more virulent. Hence, the model proposed here could help understand the transition from a moderate state of depression or even a simple negative mood to more severe phases.

Figure 4.2 Simulated contagion process



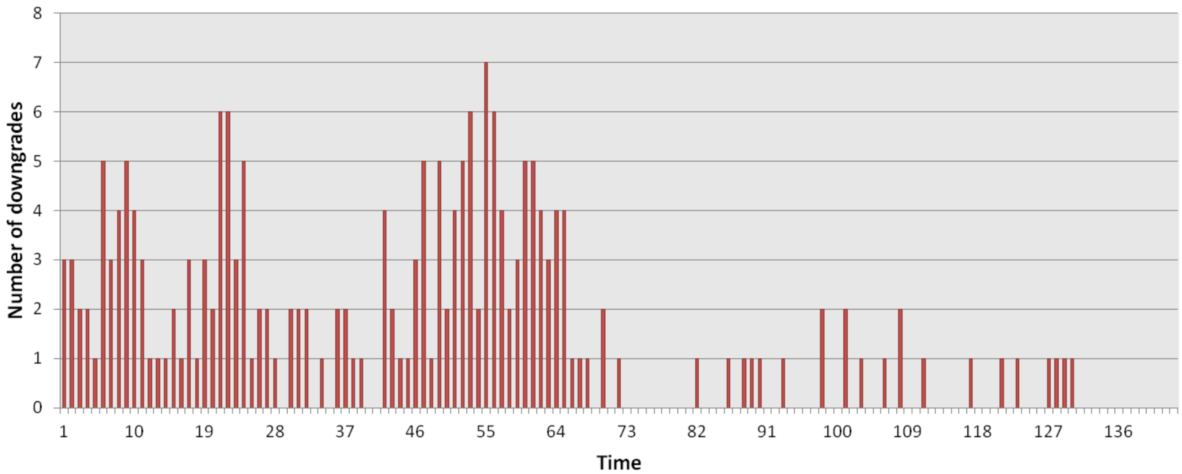
Notes: Given the initial vector of ratings (2012 Q4), this experiment simulates the contagion process showing the evolutions of the ratings for each country.

Fig. 4.2 illustrates the computer simulation on the Eurozone TPI network with the fourth quarter 2012 S&P's credit ratings as rat_0. We only use the first seventeen S&P's ratings, grouping together ratings below or equal to CCC, and translating them into a numerical scale with the highest rating defined as $AAA = 17$ and the lowest as $CCC = 1$. In this simulation, $\omega = 1$ and $\boldsymbol{\theta}$ is reported in the first column of Table 30 below.

As expected, Greece, Portugal and Cyprus are the first sovereigns to default, whereas the Netherlands, Finland, and Luxembourg are the last ones. Note that "Time" here has no specific length and should simply be interpreted as regular time-intervals. The total number of time-intervals depends on the values of $\boldsymbol{\theta}$ and ω. In particular, setting a higher ω speeds up the overall contagion process, whereas a smaller ω gives a more precise ranking of country default times.

Fig. 4.2 also illustrates an interesting network phenomenon inherent to epidemics - the "tipping point" (Gladwell [81]). The main idea is that changes often happen very quickly and unexpectedly as some threshold is reached. Here "threshold" may be interpreted as a country-specific exposure level after which its default process gains momentum. As an example, consider the evolution of Germany's credit-rating (red). It remains above AA (15) for almost 400 periods, then suddenly plummets to CCC (1) in less than 300 time-intervals. At the network level, such tipping points lead to contagion waves and clustering of default times. In fig. 4.2, three main default clusters occur around $t = 400$ (Portugal, Cyprus, Slovakia, Spain, and Slovenia), $t = 600$ (Belgium,

Figure 4.3 Contagion waves



Notes: Given the initial vector of ratings (2012 Q4), this experiment simulates the contagion process recording the number of downgrades each period.

Italy, Ireland, France, and Germany), and soon after $t = 1,200$ (Malta, Austria, The Netherlands).

To investigate this issue further, we increase the spillovers' magnitude to $\omega = 10$, and record the number of downgrades over time for a particular *contagion* realisation. Fig. 4.3 shows the emergence of contagion waves. They happen as a set of countries becomes increasingly virulent and starts transmitting downgrade spillovers at a faster pace. With each iteration of the *spillovers* function, downgrade probabilities increase for all countries, making them more likely to reach their "thresholds". This process continues until some sovereigns start defaulting. At this point, downgrade probabilities drop and the contagion process slows down because the set of highly virulent countries leaves the system. However, since downgrade probabilities for the remaining countries are still positive, a new wave eventually emerges. In practice, this experiment suggests that in order to lower the downgrade probabilities of still relatively healthy member-states and limit spillovers, the set of most virulent countries should leave the system.

While further investigation of tipping points and contagion waves is an important direction for future research, the next step and main goal of this paper is to use the model and the computer algorithm presented in this section to derive indicators of node/country importance and vulnerability.

4.3 Systemic Importance and Vulnerability

4.3.1 Early-time properties: some analytical results

Consider again the model from Section 4.2.2 and suppose that all n financial institutions start with the best possible credit-rating A, i.e. $a_{i0} = 1$ for any i. Now, let one of the institutions be exogenously downgraded. Afterwards, in order to disentangle credit risk spillovers from exogenous economic fundamentals' deterioration, set $\frac{dx_i}{dt} = 0$ for any i and t.

Which institutions are most likely to be downgraded from A in this early period? To answer this question, note that since $c_{it} = 0$, $d_{it} = 0$, and $a_{it} + b_{it} = 1$ for any i, one only needs to focus on eq.4.4 which can be rewritten as:

$$\frac{db_i}{dt} = \omega \sum_j v_{ij} b_{jt} \theta(B)$$

because $a_{it} \to 1$ and $\theta(A) = 0$. In matrix notation:

$$\frac{d\mathbf{b}}{dt} = \omega\theta(B)\mathbf{V}\mathbf{b}$$

which is a system of differential equations that has a solution of the following form:

$$\mathbf{b}(t) = \sum_{r=1}^{n} u_r(0) \exp(\omega\theta(B)\kappa_r t)\mu_{\mathbf{r}}$$

where κ_r's and $\mu_{\mathbf{r}}$'s are respectively the eigenvalues and eigenvectors of the matrix $\mathbf{V}$ and $u_r(0)$'s are some constants. Let κ_1 and $\mu_{\mathbf{1}}$ denote the largest eigenvalue and its associated eigenvector respectively:

$$\mathbf{b}(t) \sim \exp(\omega\theta(B)\kappa_1 t)\mu_{\mathbf{1}} \tag{4.8}$$

Hence, thinking of $\mathbf{b}(t)$ as an indicator of *early-time vulnerability*, eq.4.8 shows that it will be directly proportional to $\mathbf{V}'$*s right* leading eigenvector $\boldsymbol{\mu}_{\mathbf{1}}$. In the network literature, this metric is called ***(right)* eigenvector centrality** (Newman [128]).

A closely-related indicator can be constructed to gauge *early-time importance*. The trick here consists in taking the transpose of the adjacency matrix: $\mathbf{W} = \mathbf{V}^{\mathbf{T}}$ so that $w_{ij} = v_{ij}^T = v_{ji}$. Now row i of $\mathbf{W}$ gives the distribution of all network members' investment shares into i. Let $\sigma(t)$ denote the vector of *early-time importance*. We want an indicator that ranks higher those countries, whose major creditors are themselves more important. The differential equation of the form:

$$\frac{d\sigma_i}{dt} = \eta \sum_j v_{ji}\sigma_{jt}$$

shows that the importance of node i is indeed increasing in the importance of its creditors σ_{jt} and the proportion that they invest into i : v_{ji}. Rewriting in matrix notation and performing the same exercise as above one gets:

$$\sigma(t) \sim \exp(\eta\lambda_1 t)\pi_{\mathbf{1}}$$

where λ_1 and $\pi_{\mathbf{1}}$ are the leading *left* eigenvalue and eigenvector of the matrix $\mathbf{V}$. Hence the vector of early-time importance will be proportional to the ***left* eigenvector centrality**.

4.3.2 Late-time properties: simulation-based systemic indicators

One of the main goals of this paper is to measure *systemic* importance and vulnerability. To build indicators that take into account every possible feedback mechanism generated by the network structure, we need to switch from the early-time when the contagion process just sets off, to the time when the countries have traversed their default processes almost completely.

Conceptually, the experiment remains the same: all n institutions start with the best possible credit-rating A, and one of the institutions is exogenously downgraded. Nevertheless, the key difference is that instead of asking which institutions are most likely to be downgraded to rating B first, we now ask how long will it take for institution i to default? How does this time compare to the time needed for other institutions to default? What if we pick and downgrade another institution first?

A bit of imagination leads to define the following two possible indicators of systemic importance:

- *All-default time*: T_d^i gives the time when all of the nodes in the network default after the initial exogenous downgrade of node i.

- *First-default time:* F_d^i gives the time of the first default in the network after the initial exogenous downgrade of node i.

Intuitively, the initial downgrade of a systemically more important country should lead to quicker First- and All- default times (smaller F_d^i and T_d^i).

Table 27: Systemic Importance

	Rankings			Indices		
	LeftEig	Fd	Td	LeftEig	Fd	Td
Germany	1	2	1	0.469	17.535	26.652
France	3	1	2	0.456	17.527	26.792
Luxembourg	2	3	3	0.465	17.555	26.799
Netherlands	4	4	4	0.353	17.961	27.197
Italy	5	5	5	0.320	18.019	27.431
Ireland	6	6	6	0.233	18.319	27.577
Spain	7	7	7	0.225	18.543	28.109
Austria	8	8	8	0.102	20.548	29.987
Belgium	9	9	9	0.090	21.085	30.776
Portugal	10	10	10	0.058	22.250	31.840
Greece	12	11	11	0.022	22.767	33.724
Finland	11	12	12	0.036	24.862	34.559
Cyprus	15	13	13	0.003	35.323	46.868
Slovenia	13	14	14	0.005	43.168	53.474
Slovakia	14	15	15	0.004	47.989	58.222
Estonia	17	16	16	0.000	369.900	379.521
Malta	16	17	17	0.001	429.148	438.018

Note: LeftEig = Left eigenvector centrality, Fd = First default time, Td = All default time

In the same spirit, the default time of node i after the initial downgrade of node j - V_i^j - identifies i's systemic vulnerability to j. If i is more vulnerable to l than to h, we should observe: $V_i^l < V_i^h$. A summary indicator is i's average *systemic vulnerability*:

$$Vul_i = \frac{1}{n}\sum_j V_i^j$$

4.3.3 Application to the Eurozone sovereign debt crisis

To see how these different analytical and simulation-based indicators relate to each other, Tables 27 and 28 report the computed rankings and indicators for all 17 member-states using the TPI shares as the adjacency matrix (Table 26). Countries are arranged according to All-default (Td) and Vulnerability (Vul) rankings respectively.

Germany alternates with France in the role of the most systemically important Eurozone country depending on the indicator used. Although Luxembourg is a small country, it is consistently ranked second/third because other EZ members invest substantial shares into it: between 2.4% (Cyprus) and 26.2% (Italy). Remarkably, many countries (e.g. the Netherlands, Italy, Ireland, Spain) keep exactly the same positions in all importance rankings. In terms of vulnerability, the picture is less clear-cut. Slovenia

Table 28: Systemic Vulnerability

	Rankings		Indices	
	RightEig	Vul	RightEig	Vul
Slovakia	3	1	0.308	70.149
Slovenia	1	2	0.310	70.238
Estonia	5	3	0.288	70.352
Belgium	2	4	0.309	70.432
Spain	4	5	0.296	70.496
Portugal	6	6	0.287	70.518
Austria	7	7	0.274	70.549
France	9	8	0.265	70.637
Germany	10	9	0.262	70.652
Italy	8	10	0.273	70.667
Netherlands	11	11	0.213	71.136
Cyprus	15	12	0.135	71.293
Finland	12	13	0.178	72.106
Luxembourg	13	14	0.178	72.198
Malta	14	15	0.136	74.018
Ireland	16	16	0.135	75.959
Greece	17	17	0.093	77.599

Note: RightEig = Right eigenvector centrality, Vul = Average vulnerability

and Slovakia are ranked highest by different indicators. Greece is ranked lowest probably because most of its investment flows into the UK with only 26.1% remaining in the EZ. The interpretation of all the results is necessarily limited and future research should test the model on larger datasets.

To check whether these results depend on the specific values assumed for $\boldsymbol{\theta}$ and ω in the computer simulations, Table 29 contains the results from two additional simulations: with a different $\boldsymbol{\theta}$ (simulation 2), and a different ω (simulation 3). We arrange countries alphabetically and report all three simulation-based indicators. The parametrization of all simulations is summarized in Table 30.

Rankings remain almost identical. Indeed as Table 31 shows, the ranking correlations of *All-default times*, *First-default times*, and *Vulnerability* indicators across all three simulations are almost perfect.

Table 29: Robustness simulations

	Simulation 2 (different θ)						Simulation 3 (different ω)					
	Fd2		Td2		Vul2		Fd3		Td3		Vul3	
	Index	R	Index	R	Index	R	Index	R	Index	R	Index	R
Austria	74.290	8	86.760	8	617.601	7	27.339	8	56.691	8	128.670	8
Belgium	81.050	9	94.200	9	617.332	3	28.277	9	57.834	9	127.504	1
Cyprus	232.962	13	253.988	13	618.879	12	62.441	13	90.615	13	132.343	11
Estonia	4231.200	16	4243.600	16	617.475	5	714.407	16	744.862	16	127.866	4
Finland	126.770	12	139.011	12	620.943	14	36.237	11	65.681	12	134.305	13
France	39.512	2	50.385	2	617.674	8	20.961	1	50.164	1	129.066	10
Germany	39.140	1	49.989	1	617.777	9	21.018	3	50.758	2	128.825	9
Greece	83.683	10	104.333	10	624.523	17	37.772	12	64.946	11	150.784	16
Ireland	49.159	6	62.703	6	622.238	16	22.818	6	52.158	6	151.436	17
Italy	44.847	5	57.869	5	617.872	10	22.062	5	51.954	4	128.310	7
Luxembourg	39.627	3	50.399	3	619.867	13	20.974	2	51.633	3	138.959	14
Malta	4525.400	17	4537.400	17	621.473	15	879.368	17	909.022	17	139.594	15
Netherlands	43.846	4	56.643	4	618.813	11	21.790	4	52.097	5	132.440	12
Portugal	93.860	11	107.579	11	617.499	6	30.652	10	59.459	10	127.976	5
Slovakia	364.315	15	377.453	15	616.502	1	83.556	15	115.409	15	127.579	2
Slovenia	301.904	14	315.323	14	616.656	2	74.789	14	106.016	14	127.666	3
Spain	51.559	7	64.515	7	617.398	4	23.247	7	52.769	7	127.978	6

Note: R = Ranking

Table 30: Simulations parametrization

	Simulation 1	Simulation 2	Simulation 3	Positive contagion
Shock ω	100	100	50	
No. simulations	10,000	10,000	10,000	
Credit-rating	Transmission rates			
CCC	0.2958	0.1372	0.2958	0
B-	0.1479	0.1143	0.1479	0.0007
B	0.0986	0.1067	0.0986	0.0027
B+	0.0739	0.0991	0.0739	0.0060
BB-	0.0592	0.0915	0.0592	0.0107
BB	0.0493	0.0838	0.0493	0.0167
BB+	0.0423	0.0762	0.0423	0.0241
BBB-	0.0370	0.0686	0.0370	0.0328
BBB	0.0329	0.0610	0.0329	0.0428
BBB+	0.0296	0.0457	0.0296	0.0541
A-	0.0269	0.0381	0.0269	0.0668
A	0.0246	0.0305	0.0246	0.0809
A+	0.0227	0.0229	0.0227	0.0963
AA-	0.0211	0.0152	0.0211	0.1130
AA	0.0197	0.0076	0.0197	0.1310
AA+	0.0185	0.0015	0.0185	0.1504
AAA	0	0	0	0.1711

Note: For Malta and Estonia the no. of simulations is 1,000

Table 31 also sheds more light on the relationships between different indicators. There is significant positive correlation *within* the two sets of importance and vulnerability rankings, but negative correlation *between* them (bottom-left part). The positive *within* correlations suggests that the indicators capture indeed the same node characteristic (either importance or vulnerability). The difference in rankings reflects the presence of endogenous risk generated by the network structure itself. The negative *between* correlations shows that the sets of most important and most vulnerable countries do not overlap. Such a situation may lead to moral hazard problems if the most systemically important players do not bear the full network costs of idiosyncratic shocks affecting them, i.e. they do not internalize the negative externality generated on the most vulnerable countries, and take on more risks than would be socially desirable.

Table 31: Ranking correlations for Importance and Vulnerability indicators

	LEig	Fd1	Td1	Fd2	Td2	Fd3	Td3	REig	Vul1	Vul2	Vul3
LEig	1										
Fd1	0.980	1									
Td1	0.985	0.998	1								
Fd2	0.980	0.995	0.998	1							
Td2	0.980	0.995	0.998	1	1						
Fd3	0.980	0.995	0.990	0.985	0.985	1					
Td3	0.978	0.998	0.995	0.993	0.993	0.993	1				
REig	-0.115	-0.157	-0.159	-0.186	-0.186	-0.152	-0.149	1			
Vul1	-0.199	-0.213	-0.216	-0.243	-0.243	-0.216	-0.211	0.961	1		
Vul2	-0.132	-0.152	-0.154	-0.181	-0.181	-0.154	-0.149	0.968	0.990	1	
Vul3	-0.208	-0.223	-0.221	-0.248	-0.248	-0.228	-0.211	0.953	0.961	0.961	1

4.4 Positive Contagion

"We spoke a lot about contagion when things go poorly but I believe there is a positive contagion when things go well. "

Mario Draghi, summer 2012

4.4.1 A world of two contagions

In the model analysed up to now, the only possible transition for countries was downwards. For instance, in the simple model with only four ratings from Section 4.2.2, we had:

$$A \rightarrow B \rightarrow C \rightarrow D$$

In reality, however, countries can also be upgraded; and as Draghi's quotation suggests, the sign of contagion is probably determined on a daily basis by market news and sentiments. An announcement like the "Draghi Put" or a successful summit might lead to a round of positive contagion, whereas a failed government bonds auction or the publication of exorbitant youth unemployment rates may induce negative contagion. In this more realistic world of two contagions - positive and negative - the transition pattern becomes:

$$A \leftrightarrows B \leftrightarrows C \rightarrow D$$

Akin to negative credit spillovers, positive ones occur if institution/sovereign i's initial upgrade leads to an increase in the upgrade probabilities of its creditors. This section extends the basic default model with only four credit ratings to allow for this possibility.

For simplicity, let us ignore the exogenous factors that could lead to upgrades and downgrades and concentrate on credit changes that happen because of credit spillovers. Let $\xi(cr_{it}) \in [0;1]$ be the rate at which institution i with credit-rating cr_{it} transmits a positive spillover to its creditors. In this case, the positive transmission rates vector $\boldsymbol{\xi}$ shall be increasing in cr, reflecting the fact that it becomes easier for an institution to transmit positive spillovers as its rating rises. Since a defaulted organisation cannot transmit positive spillovers, we can further simplify the vector as: $\xi(D) = 0$, and $\xi(cr) \in (0;1]$ for any $cr \neq D$.

For institution i to be upgraded from B to A thanks to a positive credit spillover from one of its debtors j, it must itself have rating B at the outset - which happens with probability b_{it}, one of its debtors must have rating $cr_{jt} \in\{A,\ B,\ C\}$ - which happens with probabilities $\psi_{jt} = \{a_{jt},\ b_{jt},\ c_{jt}\}$ respectively - and transmit the positive spillover at rate $\xi(cr_{jt})$.

In a world of two contagions, the differential equation that describes the probability that i has rating A at any point in time is therefore:

$$\frac{da_i}{dt} = -\omega \sum_j v_{ij} \left\{a_{it} z_{jt}\right\} \theta(cr_{jt}) + \phi \sum_j v_{ij} \left\{b_{it} \psi_{jt}\right\} \xi(cr_{jt}) \tag{4.9}$$

where $\phi \in \mathbb{R}^+$ captures the magnitude of the positive spillovers. Here we assume that both contagions propagate through the same network - the network of investment shares with adjacency matrix $\mathbf{V}$. A more complex and realistic model would allow positive and negative contagions to spread through multiple and/or different networks.

Similarly, the probability that i has rating B increases in i's probability of being downgraded from A to B, decreases in i's probability of being downgraded from B to C, decreases in i's probability of being upgraded from B to A, and increases in i's probability of being upgraded from C to B.

$$\begin{aligned}\frac{db_i}{dt} =& \,\omega \sum_j v_{ij} \left\{a_{it} z_{jt}\right\} \theta(cr_{jt}) - \omega \sum_j v_{ij} \left\{b_{it} z_{jt}\right\} \theta(cr_{jt}) \\ & - \phi \sum_j v_{ij} \left\{b_{it} \psi_{jt}\right\} \xi(cr_{jt}) + \phi \sum_j v_{ij} \left\{c_{it} \psi_{jt}\right\} \xi(cr_{jt})\end{aligned} \tag{4.10}$$

Since once defaulted, an institution can no longer be upgraded, the probability that i has rating C increases in i's probability of being downgraded from B to C, decreases in i's probability of being downgraded from C to D, and decreases in i's probability of being upgraded from C to B:

$$\frac{dc_i}{dt} = \omega \sum_j v_{ij} \left\{b_{it} z_{jt}\right\} \theta(cr_{jt}) - \omega \sum_j v_{ij} \left\{c_{it} z_{jt}\right\} \theta(cr_{jt}) - \phi \sum_j v_{ij} \left\{c_{it} \psi_{jt}\right\} \xi(cr_{jt}) \quad (4.11)$$

Finally, the probability of defaulting remains as before:

$$\frac{dd_i}{dt} = \omega \sum_j v_{ij} \left\{c_{it} z_{jt}\right\} \theta(cr_{jt}) \quad (4.12)$$

Note that this is a more general version of the model presented in Section (4.2.2) where ϕ - the magnitude of positive spillovers - was equal to zero, i.e. the positive contagion channel was shut down.[58]

4.4.2 Can positive contagion save the Eurozone?

"Where's your positive contagion now, Mr. Draghi?"

J. Warner, *The Telegraph* (05/02/2013)

To analyse this more complex model through computer simulations we can define:

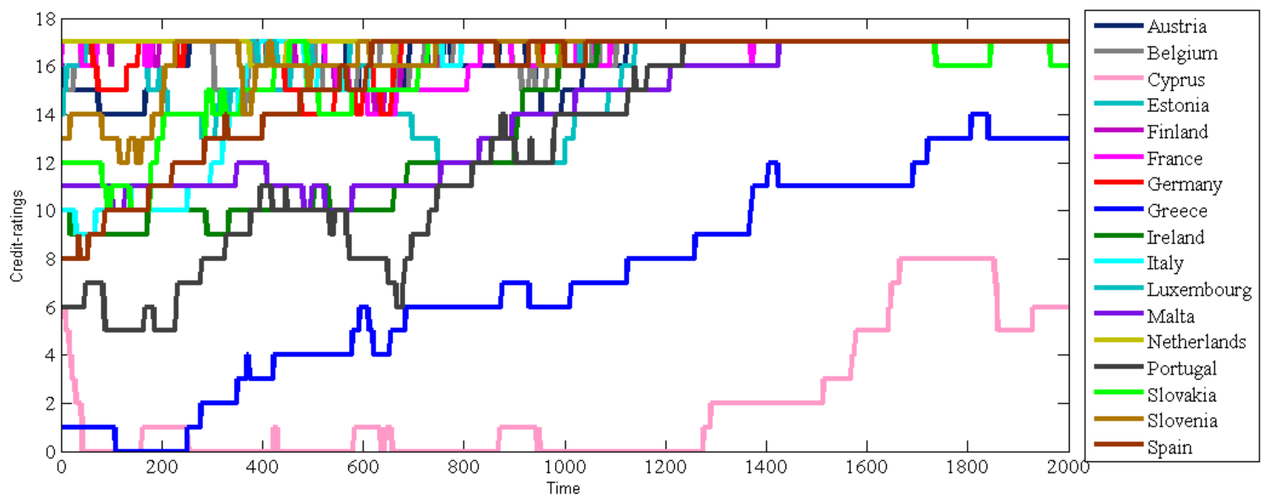
$$\mathbf{Pr}(i\ upgraded) = \phi \sum_j v_{ij} \xi(cr_j) \quad (4.13)$$

The new extended *spillovers'* function in the computer algorithm now associates to each country i two transmission rates $\theta(cr_i)$ and $\xi(cr_i)$. After computing country-specific downgrade and upgrade probabilities using equations 4.7 and 4.13, some countries are stochastically downgraded and upgraded producing a new vector of ratings.

Whether positive contagion ultimately dominates its negative counterpart will depend both on the transmission rates vectors $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, and the respective spillover mag-

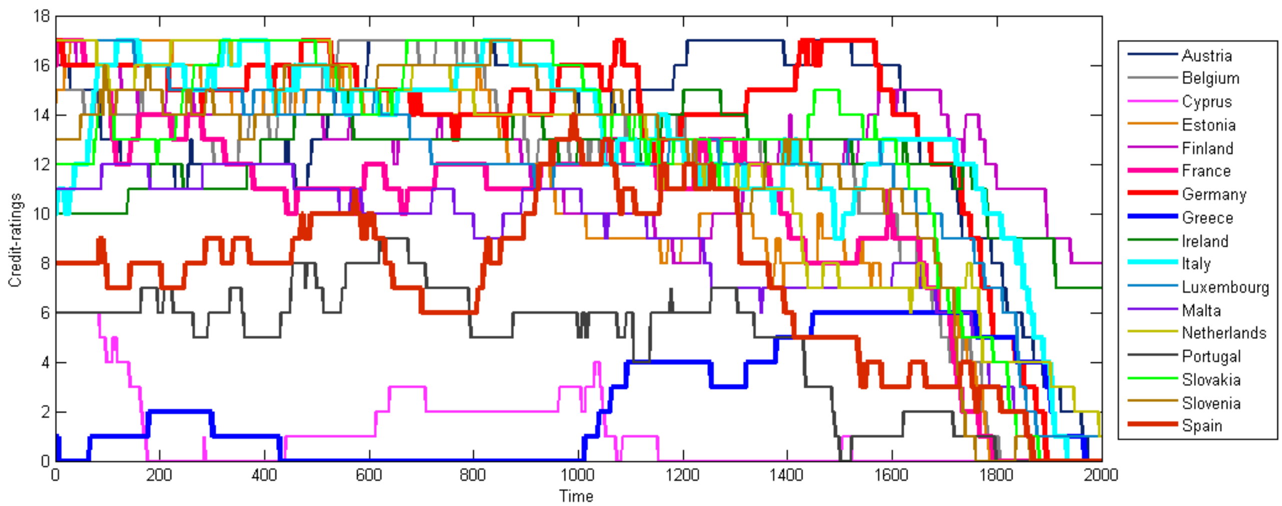
[58] Following our discussion in footnote (57) about how this multi-stage epidemic model might be used in the context of contagious mental illnesses, note that this extended version with both positive and negative contagions could allow researchers to investigate in a more systematic way how depressed people's well-being is enhanced or eroded by positive and negative social interactions." (Kashdan & Steger [108]). Different networks, reflecting the type and intensity of the interaction between people, could be used to propagate the latter.

Figure 4.4 Contagion dynamics (1)



Notes: $\omega = 1$; $\phi = 0.3$

Figure 4.5 Contagion dynamics (2)



Notes: $\omega = 1$; $\phi = 0.3$

nitudes ω and ϕ. While future research shall investigate this important issue in a more details, Figures 4.4 and 4.5 illustrate how, for given $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ and ω, lowering the positive contagion magnitude ϕ from 0.3 to 0.2 changes the scenario from one where most of the countries finish with rating AAA, to a more morose one, with complex spillover dynamics.[59]

This suggests that unless Mr. Draghi manages to make "things go well [enough]", we will have difficulty in noticing positive contagion.

4.5 Conclusion & Future Research

> *"The deadliest aspect of the Eurozone crisis is the tripwire linking the riskiness of banks and governments."*
>
> Acharya et al. [2]

Most of the literature on financial contagion in networks has concentrated on analysing how an initial default of a bank or country triggers a cascade of further failures. The Eurozone sovereign debt crisis however has demonstrated that although *defaults* are likely to be prevented, *credit-rating downgrades* occur rather often. Since credit-ratings are one of the key drivers of investment decisions, a model explaining credit risk spillovers could help understand the observed reallocation of investors' portfolios in response to changes in the creditworthiness of interlinked countries or banks. To identify such credit risk spillovers, this paper models the process of default as a multi-stage disease with each credit-rating corresponding to a different infection phase. We use the model to develop indicators of systemic importance and vulnerability by investigating how the initial exogenous change in the creditworthiness of one of the members of the financial network impacts the creditworthiness of all the other ones.

The illustration of the model in the context of the Eurozone sovereign debt crisis yields interesting and intuitive results. For example, we find that France and Germany occupy the highest positions in the systemic importance hierarchy. However, the interpretation of all the results is necessarily limited by the small dataset. Future research should include more countries and test the model on interbank data, as well as investigate more carefully such phenomena as tipping points and contagion waves. The above quote from Acharya et al. [2] implies that governments and banks are closely interlinked. Hence further research should not only analyse the networks of banks and

[59] The positive transmission-rates vector used is reported in Table 30.

governments separately, but also investigate how an initial shock in one of the networks could potentially engender contagion in the other one, or even change its structure.

The literature has focused almost exclusively on negative contagion. However, policymakers seem to be aware that the same endogenous feedback mechanisms that yield negative financial contagion, could in principle be used to activate positive spillovers and shift investors' sentiments. Unfortunately, the extension of the model to include such positive spillovers shows that even though positive contagion could be an attractive solution to the Eurozone sovereign debt crisis, the main policy question remains how to generate it permanently, halting its negative counterpart. Another task for further research is therefore to examine how the ratio of positive to negative spillover magnitudes determines the overall sign of contagion.

References

[1] Daron Acemoglu, Simon Johnson, and James A. Robinson. 'Institutions as the fundamental cause of long-run growth'. *Handbook of Economic Growth, Volume IA. Edited by Philippe Aghion and Steven N. Durlauf*, 2005.

[2] Viral V. Acharya, Itamar Drechsler, and Philipp Schnabl. 'A pyrrhic victory? Bank bailouts and sovereign credit risk'. *Working paper, NYU-Stern, CEPR Discussion Paper 8679*, 2010.

[3] A. Afonso. 'Understanding the determinants of sovereign debt ratings: Evidence of the two leading agencies'. *Journal of Economics and Finance*, 27:56–74, 2003.

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

[5] Philippe Aghion and Ufuk Akcigit. 'Innovation and Growth: The Schumpeterian Perspective'. *COEURE Coordination Action*, October, 2015.

[6] Philippe Aghion, Gilbert Cette, and Élie Cohen. *'Changer de modèle'*. Odile Jacob, 2014.

[7] Mark Aguiar and Gita Gopinath. 'Emerging Market Business Cycles: The Cycle is the Trend'. *Journal of Political Economy, 115(1)*, 2007.

[8] F. Allen and A. Babus. 'Networks in finance'. *The Network Challenge*, edited by P. Kleindorfer and J. Wind, Wharton School Publishing:367–382, 2009.

[9] J. Allen and R. van der Velden. 'Educational mismatches versus skill mismatches: Effects on wages, job satisfaction, and on-the-job search'. *Oxford Economic Papers*, 3, 2001.

[10] Arthur Allignol, Jan Beyersmann, and Martin Schumacher. *'Competing Risks and Multistate Models with R'*. Springer Science Business Media, Use R, 2012.

[11] Paul D. Allison. 'Discrete-time methods for the analysis of event histories'. *Sociological Methodology, Vol. 13*, 1982.

[12] M. Dell' Amico and P. Toth. 'Algorithms and codes for dense assignment problems: The state of the art'. *Discrete Applied Mathematics*, 100, 2000.

[13] R. Arezki, B. Candelon, and A.N.R. Sy. 'Sovereign rating news and financial markets spillovers: Evidence from the european debt crisis'. *IMF Working Paper*, 2011.

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

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

[16] Carlo Azzarri, Alberto Zezza, Benjamin Davis, Katia Covarrubias, Luca Tasciotti, and Gustavo Anriquez. 'The impact of rising food prices on the poor'. *ESA Working Paper No. 08-07*, August, 2008.

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

[18] Jushan Bai and Pierre Perron. 'Estimating and testing linear models with multiple structural change'. *Econometrica*, 66:47–78, 1998.

[19] Jushan Bai and Pierre Perron. 'Computation and analysis of multiple structural change models'. *Journal of Applied Economics*, 18:1–22, 2003.

[20] Jushan Bai and Pierre Perron. 'Multiple structural change models: A simulation analysis'. *Econometric Theory and Practice: Frontiers of Analysis and Applied Research. Cambridge University Press*, pages 212–237, 2006.

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

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

[23] Dan Ben-David and David H. Papell. 'Slowdowns and meltdowns: Postwar growth evidence from 74 countries'. *Review of Economics and Statistics*, 80(4), 1998.

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

[25] Andrew Berg, Jonathan D. Ostry, and Jeromin Zettelmeyer. 'What makes growth sustained?'. *Journal of Development Economics*, 98:149–166, 2012.

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

[27] Emawtee Bissoondoyal-Bheenick and Sirimon Treepongkaruna. 'An analysis of the determinant of bank ratings: Comparison across ratings agencies.'. *Australian Journal of Management 36:405*, 2011.

[28] 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.

[29] 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.

[30] 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.

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

[32] G. M. Caporale, A. Serguieva, and H. Wu. 'Financial contagion: evolutionary optimization of a multinational agent-based model'. *Intelligent Systems in Accounting, Finance and Management. Intell. Sys. Acc. Fin. Mgmt*, 16:111–125, 2008.

[33] P. Cappelli. *Why Good People Can't Get Jobs: The Skills Gap and What Companies Can Do About It.* Wharton Digital Press, 2012.

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

[35] CBI/KPMG. 'London business survey'. 2014.

[36] C. Célérier and B. Vallée. 'The returns to talent and the finance wage premium'. *Working paper*, 2015.

[37] S. Cheung. ‘Provincial credit ratings in Canada: An ordered probit analysis’. *Bank of Canada, Working Paper 96-6*, 1996.

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

[39] Yu-Wei Chiu. *‘Machine Learning with R Cookbook’*. Packt Publishing, March, 2015.

[40] European Commission. ‘Employers’ perception of graduate employability’. *Flash Eurobarometer No. 304*, 2010.

[41] G. Corsetti, P. Pesenti, and N. Roubini. ‘Paper tigers? A model of the Asian crisis’. *European Economic Review*, 43, 1999.

[42] D. R. Cox. ‘Regression models and life-tables’. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2*, 1972.

[43] V.P. Crawford and E.M. Knoer. ‘Job matching with heterogeneous firms and workers’. *Econometrica*, 49, No.2, 1981.

[44] David Deming and Lisa B. Kahn. ‘Firm heterogeneity in skill demands’. June 2016.

[45] N. Demiris, T. Kypraios, and L.V. Smith. ‘On the epidemic of financial crises’. *JEL Classification: C32, E17, F37, F42*, 2012.

[46] D. W. Diamond and P.H. Dybvig. ‘Bank runs, deposit insurance, and liquidity’. *Journal of Political Economy*, 91:401–419, 1983.

[47] Carlos F. Diaz-Alejandro. ‘Latin American debt: I don’t think we are in Kansas anymore’. *Brookings Papers on Economic Activity, 2*, 1984.

[48] J. Dimon and M. Seltzer. ‘Closing the skills gap’. *PoliticoMagazine*, January 05, 2014.

[49] William Easterly, Michael Kremer, Lant Pritchett, and Lawrence H. Summers. ‘Good policy or good luck: country growth performance and temporary shocks’. *Journal of Monetary Economics, 32(3)*, pages 459–83, 1993.

[50] B. Eichengreen, A. Rose, and C. Wyplosz. ‘Contagious currency crises’. *Scandinavian Journal of Economics*, 1996.

[51] Liran Einav and Jonathan D. Levin. 'The data revolution and economic analysis.'. *NBER Working Paper No. 19035*, May 2013.

[52] H. A. Eiselt and C.-L. Sandbloml. *Operations Research: A Model-Based Approach (2nd Edition).* Springer, 2012.

[53] M. Elliott, B. Golub, and M.O. Jackson. 'Financial networks and contagion'. *American Economic Review, 104(10)*, 2014.

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

[55] M. A. Espinosa-Vega and J. Sole. 'Cross-border financial surveillance: A network perspective'. *IMF Working Paper WP/10/105*, 2010.

[56] J.D. Farmer and D. Foley. 'The economy needs agent-based modelling'. *Nature*, 460:685–686, 2009.

[57] Robert C. Feenstra, Robert Inklaar, and Marcel P. Timmer. 'The next generation of the Penn World Table'. *American Economic Review, 105(10)*, pages 3150–3182, 2015.

[58] Carmem Feijo, Lionello F. Punzo, and Marcos Tostes Lamonica. 'Brazil's economy- 1971-2005: Growth pattern and structural change'. *Center for Studies on Inequality and Development, Discussion Paper No. 15*, April 2009.

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

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

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

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

[63] UK Commission for Employment and Skills. 'Working Futures: 2014-2024'. *UKCES Evidence Report 100*, April, 2016.

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

[65] Department for Innovation, Universities, and Skills. 'The demand for Science, Technology, Engineering and Mathematics (STEM) skills '. January 2009.

[66] Office for National Statistics. 'Graduates in the labour market'. 19 November 2013.

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

[68] The European Centre for the Development of Vocational Training (Cedefop). 'Skill mismatch: Identifying priorities for future research'. *Cedefop working paper No. 3*, May 2009.

[69] John Forth and Geoff Mason. 'Do ICT skill shortages hamper firms' performance? Evidence from UK benchmarking surveys'. *National Institute of Economic and Social Research, London*, September, 2006.

[70] Jean Fourastié. *'Les Trente glorieuses ou la Révolution invisible de 1946 à 1975'*. Fayard, 1979.

[71] Philip Hans Franses, Dick van Dijk, and Richard Paap. 'Does Africa grow slower than Asia and Latin America?'. *Econometric Institute Report EI 2003-07*, February 2003.

[72] Mark W. French. 'Estimating changes in trend growth of total factor productivity: Kalman and H-P filters versus a Markov-switching framework'. *FEDS Working Paper No. 2001-44*, September 2001.

[73] 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.

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

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

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

[77] P. Gai, A. Haldane, and S. Kapadia. 'Complexity, concentration and contagion'. *Journal of Monetary Economics*, 58:453–470, 2011.

[78] D. Gale and L. Shapley. 'College admissions and the stability of marriage'. *American Mathematical Monthly*, 69, 1962.

[79] M. Galenianos and P. Kircher. 'Directed search with multiple job applications'. *Journal of economic theory*, 144 (2):445–471, 2009.

[80] S. Gerlach and F. Smets. 'Contagious speculative attacks'. *European Journal of Political Economy*, 11:5–63, 1995.

[81] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown, 2000.

[82] Patricia M. Grambsch and Terry M. Therneau. *'Modeling Survival Data: Extending the Cox Model '*. Springer, 2000.

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

[84] 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.

[85] Ted Robert Gurr, Monty G. Marshall, and Keith Jaggers. *'POLITY IV PROJECT: Political Regime Characteristics and Transitions, 1800-2015, Dataset Users' Manual'*. Center for Systemic Peace, 2016.

[86] A.G. Haldane. 'Rethinking the financial network'. *Speech delivered at the Financial Student Association, Amsterdam*, 2009.

[87] Matthew Harrison. 'Jobs and growth: the importance of engineering skills to the UK economy'. *Royal Academy of Engineering econometrics of engineering skills project*, September, 2012.

[88] 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.

[89] Ricardo Hausmann, Lant Pritchett, and Dani Rodrik. 'Growth accelerations'. *Journal of Economic Growth*, 10:303–329, 2005.

[90] Ricardo Hausmann, Francisco Rodriguez, and Rodrigo Wagner. 'Growth collapses'. *Kennedy School of Government Working Paper RWP06-046*, 2006.

[91] Hays and Oxford Economics. 'Hays global skills index'. 2014.

[92] J. Heckman and J. Scheinkman. 'The importance of bundling in a Gorman-Lancaster model of demand for characteristics: A theoretical analysis with evidence from the labor market'. *Review of Economic Studies*, 54(2):243–255, 1987.

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

[94] Matthew Higgins. 'Demography, national savings, and international capital flows'. *International Economic Review Vol.39(2)*, May, 1998.

[95] R. Hodrick and E.C. Prescott. 'Post-war U.S. business cycles: An empirical investigation'. *Discussion Paper at Northwestern University and Carnegie-Mellon University*, 1980.

[96] R. Holt, S. Sawicki, and J. Sloan. 'A theoretical review of skill shortages and skill needs'. *Evidence Report 20, UK Commission for Employment and Skills*, May 2010.

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

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

[99] V. Jaquet and M. Pechal. 'Epidemic spreading in a social network'. *ETH Zurich, Project Report*, 2009.

[100] Stephen P. Jenkins. *'Survival Analysis'*. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK, 2005.

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

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

[103] J.Janssen. 'Spatial models for virtual networks'. *Lecture Notes in Computer Science*, 6158, 2010.

[104] Thomas E. Joiner and Jennifer Katz. 'Contagion of depressive symptoms and mood: Meta-analytic review and explanations from cognitive, behavioral, and interpersonal viewpoints.'. *American Psychological Association, Clinical Psychology: Science and Practice, V6 N2*, 1999.

[105] F. Benjamin Jones and Benjamin A. Olken. 'The anatomy of start-stop growth'. *Review of Economics and Statistics 90 (3)*, pages 582–587, 2008.

[106] G. L. Kaminsky and C. M. Reinhart. 'Financial crises in Asia and Latin America: Then and now'. *American Economic Review*, 88:444–448, 1998.

[107] Sabyasachi Kar, Lant Pritchett, Selim Raihan, and Kunal Sen. 'Looking for a break: Identifying transitions in growth regimes'. *Journal of Macroeconomics*, pages 151–166, 2013.

[108] T.B. Kashdan and M.F. Steger. 'Depression and everyday social activity, belonging, and well-being.'. *Journal of Counseling Psychology, 56(2)*, 2009.

[109] Monika Kerekes. 'Analyzing patterns of economic growth: a production frontier approach'. *Discussion Paper, Free University Berlin*, February, 2008.

[110] W. O. Kermack and A. G. McKendrick. 'A contribution to the mathematical theory of epidemics'. *Proceedings of the Royal Society of London.*, 138:55–83, 1932.

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

[112] R.E. Kranton and D.F. Minehart. 'Competition for goods in buyer-seller networks'. *Review of Economic Design*, 5, 2000.

[113] P. Krugman. 'Jobs and skills and zombies'. *The New York Times*, March 30, 2014.

[114] H.W. Kuhn. ‘The Hungarian method for the assignment problem’. *Naval Research Logistics Quarterly*, 2, 1955.

[115] K.J. Lancaster. ‘A new approach to consumer theory’. *Journal of Political Economy*, 74(2):132–157, 1966.

[116] Dominique Lejeune. *‘La France des Trente Glorieuses - 1945-1974’.* ARMAND COLIN, June 2015.

[117] 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.

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

[119] ManpowerGroup. ‘The talent shortage continues’. 2014.

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

[121] P. R. Masson. ‘Multiple equilibria, Contagion and the Emerging Market Crises’. *IMF Working Paper No. 99/164*, 1999.

[122] R.M. May, S.A. Levin, and G. Sugihara. ‘Complex systems: Ecology for bankers’. *Nature*, 451:893–895, 2008.

[123] S. McGuinness. ‘Overeducation in the labour market’. *Journal of Economic Surveys*, 20, No.3, 2006.

[124] C. Melliosa and E. Paget-Blanc. ‘Which factors determine sovereign credit ratings?’. *European Journal of Finance*, 12:361–377, 2006.

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

[126] E.R. Moen. ‘Competitive search equilibrium’. *Journal of Political Economy*, 105, 1997.

[127] Ricardo Mora and Siotis Georges. ‘External factors in emerging market recoveries: An empirical investigation’. *European Economic Review 49*, 2005.

[128] M.E.J. Newman. *Networks: An Introduction.* Oxford University Press, 2012.

[129] Friedrich Wilhelm Nietzsche. *Schopenhauer as Educator.* Adrian Collins (Translator), 1873.

[130] P. Pesenti and C. Tille. 'The economics of currency crises and contagion: An introduction'. *FRBNY Economic Policy Review*, 2000.

[131] C. Pissarides. 'Equilibrium unemployment theory'. *Cambridge, Massachusetts: Basil Blackwell*, 1990.

[132] Lant Pritchett. 'Understanding patterns of economic growth: searching for hills among plateaus, mountains and plains'. *World Bank Economic Review 14(2)*, pages 221–250, 2000.

[133] Lant Pritchett and Lawrence H. Summers. 'Asiaphoria Meets Regression to the Mean'. *NBER Working Paper No. 20573*, October, 2014.

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

[135] G. Rodriguez. 'Lecture notes on generalized linear models'. 2007.

[136] Paul M. Romer and Zvi Griliches. 'Implementing a national technology strategy with self-organizing industry investment boards'. *Brookings Institution*, 1993.

[137] S. Rosen. 'Hedonic prices and implicit markets: Product differentiation in pure competition'. *Journal of Political Economy*, 82:34–55, 1974.

[138] A.E. Roth, T. Sonmez, and M. U. Unver. 'Kidney exchange'. *Quarterly Journal of Economics*, 119(2):457–488, 2004.

[139] A.E. Roth and M.A.O. Sotomayor. *Two-sided matching : A study in game-theoretic modeling and analysis.* Cambridge University Press, 1990.

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

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

[142] A.D. Roy. 'Some thoughts on the distribution of earnings'. *Oxford Economic Papers*, 3.2, 1951.

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

[144] M. Sattinger. 'Comparative advantage and the distributions of earnings and abilities'. *Econometrica*, 43, No.3, 1975.

[145] M. Sattinger. 'Qualitative mismatches'. *Foundations and Trends in Microeconomics*, 8, Nos.1-2:1–168, 2012.

[146] Franklin Serrano and Ricardo Summa. 'Aggregate demand and the slowdown of Brazilian economic growth from 2011-2014'. *Center for Economic and Policy Research*, August 2015.

[147] H. Shierholz. 'Is there really a shortage of skilled workers?'. *Restoring Shared Prosperity: A Policy Agenda from Leading Keynesian Economists, edited by Thomas I. Palley and Gustav A. Horn.*, 2013.

[148] R. Shimer. 'The assignment of workers to jobs in an economy with coordination frictions'. *Journal of Political Economy*, 113(5):996–1025, 2005.

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

[150] R.M. Solow. 'A contribution to the theory of economic growth'. *Quarterly Journal of Economics 70 (1)*, pages 65–94, 1956.

[151] Nikola Spatafora and Issouf Samake. 'Commodity price shocks and fiscal outcomes'. *IMF Working Paper WP/12/112*, May, 2012.

[152] Silvana Tenreyro. 'The Argentine crisis in retrospect'. *London School of Economics and Political Sciences*, July, 2012.

[153] C. N. Teulings and J. A. C. Vieira. 'Urban versus rural return to human capital in Portugal: A cookbook recipe for applying assignment models'. *LABOUR*, 18:265–291, 2004.

[154] J. Tinbergen. 'Some remarks on the distribution of labour incomes'. *Internat. Econ. Papers*, 1, 1951.

[155] J. Tinbergen. 'On the theory of income distribution'. *Weltwirtschaftliches Archiv*, 77(1):156–75, 1956.

[156] Christian Upper. ‘Contagion due to interbank credit exposures: What do we know, why do we know it, and what should we know?’. *Bank for International Settlements*, 2006.

[157] Hal R. Varian. ‘Big Data: New Tricks for Econometrics.’. *Journal of Economic Perspectives*, 28:3–28, Spring 2014.

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

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

[160] Kenneth A. Wilson. *‘CROSS-NATIONAL TIME-SERIES DATA ARCHIVE: Users’ Manual’*. 2016.

[161] Rob Wilson. ‘The demand for STEM graduates: some benchmark projections’. *Council for Industry and Higher Education, London*, January, 2009.

[162] Linda Yueh. ‘Thirty years of legal and economic reform in China: growth, institutions, and laws’. *Journal of Chinese Economic and Business Studies, 8:2*, pages 117–132, 2010.

[163] Achim Zeileis, Christian Kleiber, Walter Kramer, and Kurt Hornik. ‘Testing and dating of structural changes in practice’. *Computational Statistics and Data Analysis 44*, 2003.

[164] Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. ‘strucchange: An R package for testing for structural change in linear regression models’.

[165] Xiaodong Zhu. ‘Understanding China’s growth: Past, present, and future’. *Journal of Economic Perspectives, Vol. 26(4)*, 2012.

[166] Jean-Pierre Zigrand. ‘Systems and systemic risk in finance and economics’. *SRC Special Paper No. 1*, 2014.