JUSTUS LIEBIG UNIVERSITY GIESSEN

DISSERTATION

Essays on the Application of Statistical Learning in Empirical Economic Research

Author:
Julian Oliver Dörr

Supervisory Committee:
Prof. Dr. Peter Winker
Prof. Dr. Christian Aßmann

Submitted in fulfillment of the requirements for the degree of

DOCTOR RERUM POLITICARUM (Dr. rer. pol.)

in the

Faculty of Economics and Business Studies
Department of Statistics and Econometrics

May 31, 2022

Declaration of Authorship

I, Julian Oliver Dörr, declare that this thesis titled, "Essays on the Application of Statistical Learning in Empirical Economic Research" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Munich, 31/05/22 J. Doerr

Place, date Signature

Eigenständigkeitserklärung

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Munich, 31/05/22 J. Doerr

Ort, Datum Unterschrift

Acknowledgements

I would like to thank the ZEW - Leibniz Centre for European Economic Research and the Justus Liebig University Giessen for offering me a stimulating research environment, critical infrastructure and funding that were integral for the completion of my dissertation. Special thanks goes to my supervisor Peter Winker. From the very beginning he approached me open-mindedly, pointed me into useful research directions and encouraged me in critical moments. Without his continued support throughout this project, I would not have come to writing these lines. I would also like to extend my gratitude to my second supervisor Christian Aßmann, from whom I have learned a lot during my studies and who always took the time for discussions - also on a personal level. Many thanks also to the members of my dissertation committee - not only for their time, but also for their intellectual contributions to my work. I am very grateful to my colleagues at the department of Economics of Innovation and Industrial Dynamics at ZEW. I will always have fond memories of your respectful interaction with each other and the great collegial cohesion as well as the intellectually enriching discussions. Especially helpful to me during this time was Georg Licht who, to me, was more than just a department head but a true mentor. I would like to thank my co-authors Simona Murmann, Jan Kinne and David Lenz for the great teamwork. Simona, the many calls and discussions we had were the best time during my thesis. I highly appreciate your unwavering support. Jan, your positive attitude and doer mentality were always encouraging. David, million thanks for assisting me with my many methodological and administrative questions. I am also grateful to Tobias Weih for insightful discussions and all the support throughout this project. Finally, but not least, I want to deeply thank my family and friends for all the unconditional support and understanding during this intense time.

Contents

List of Figures

List of Tables

Preface

This thesis was written between June 2020 and May 2022 as external Ph.D. candidate at the Chair of Statistics and Econometrics at the Justus Liebig University Giessen. At the time of writing, the author was employed as researcher at ZEW - Leibniz Centre for European Economic Research in the department of Economics of Innovation and Industrial Dynamics. The thesis is submitted in partial fulfillment of a Ph.D. degree. It consists of four chapters and contains three separate research articles.

Chapter 1 gives an introduction to the rise of statistical learning methods in economic research. Inspired by the significant advances of these methods in recent years and their potential to open new avenues in economic research, this dissertation contributes in the form of three papers that use statistical learning methods to answer open research questions about the behavior of firms and economic policymakers under dynamic market conditions. Chapter 1 briefly introduces these research articles and provides a brief discussion on how statistical learning was used to answer the papers' scope of research. All articles are reprinted in the Appendix. Chapter 2 consists of two articles whose common linkage is the dynamics triggered by the unprecedented economic shock of COVID-19. Statistical learning methods were used to guide policymakers in their response measures and to evaluate the effect of policy interventions on firm closure dynamics. The second article in Chapter 2 already foreshadows the following work in Chapter 3, as the methodological focus is on the analysis of textual data and its application as a policy tool. The third paper presented in Chapter 3 introduces a novel text modeling approach to map technologies to business models, opening up a new possibility to evaluate technology profiles of market entrants.

Article 1 (Chapter 2) empirically analyzes whether government support for ailing firms in the wake of the first COVID-19 induced lockdown led to a backlog of corporate insolvencies and, should this be the case, whether this backlog is disproportionally characterized by firms that were already in distress before COVID-19 hit. This might hint at the unwanted side-effect of interfering with Schumpeter's natural market cleansing dynamics. For the estimation strategy, the paper makes use of a matching approach that builds on the k Nearest Neighbor (kNN) algorithm. This method of supervised learning allows for a comparison of companies with similar characteristics that have experienced almost the same updates to their credit ratings under different policy regimes, i.e., before and after the pandemic outbreak. The paper is a joint work with Georg Licht and Simona Murmann. My share in this article is 60%.

Article 2 (Chapter 2) proposes a data framework that allows to assess the impact of an unforeseen economic shock at firm level and at near real-time. It shows that different sources of impact data can be integrated into a policy tool to overcome information deficits that policymakers typically face in highly dynamic situations such as at the beginning of the COVID-19 pandemic. Moreover, the framework shows that businesses' communication patterns concerning the pandemic serve to forecast deterioration in their financial standing over the course of the crisis. The article is co-authored with Jan Kinne, David Lenz, Georg Licht and Peter Winker. My own share is 60%.

In Article 3 (Chapter 3), I develop an approach to map technologies to business models based on a topic model architecture and text embedding models. In the paper, I show how patent texts and business descriptions can be transferred in a common vector space to measure companies' technological orientation. The theoretical contribution of the paper is concerned with the role of market entrants in the diffusion and development of environmentally sound technologies. The article is single authored.

Chapter 4 concludes and suggests avenues for further research that could benefit from the work presented in this thesis.

Chapter 1

Introduction

Digitization has increasingly made its way into our everyday lives. As a result, we are leaving digital traces on a daily basis. This applies not only to us as consumers, but also to all other economic actors such as companies, policymakers and entrepreneurs. The increased digital recording of economic activity has made it possible to investigate economic relationships for which the data basis was previously fragmentary or even lacking (Einav & Levin, 2014). Whether it is price movements on Amazon, bidding processes on eBay, communication patterns on social media, product descriptions on corporate websites or night light data from satellite images, the digitized content of economic transactions, interactions and (re-)actions provides economists with an almost inexhaustible information reservoir for tracking economic activity and trying to understand the consequences of different economic policies.

However, increasing digitization has also added complexity to economically-relevant information sources, both in terms of scale and structure. In the era of "big data", technical hurdles to handle large sets of research data pose one source of complexity - accessing and processing digital archives covering terabytes of corporate website data, for example, requires technical knowledge of parallel computing. Besides the increase in scale, the nature of modern data that promises great leverage for economic research has also changed (Mullainathan & Spiess, 2017). In fact, data complexity is increasingly determined by the unstructured nature of research data, as in the case of image or language data. Turning these complex forms of data into meaningful economic indicators that can be fed into regressions, requires skills and methods that typically go beyond the econometrician's standard toolkit. Early research in this field used simplified extraction methods to retrieve economically meaningful signals from such unstructured data.

Henderson et al. (2012), for example, use nighttime lights from satellite data to construct a regional measure of economic performance that is independent from national borders. Their findings are based on satellite images with a low spatial resolution of 1 km and a predefined six-bit digital number reflecting the grids luminosity. Today, satellite data is available on a much higher scale with a resolution of up to

0.5 m and, methodologically, state-of-the-art convolutional neural networks can successfully detect more complex patterns from images than just luminosity. Language data is another unstructured information source that has found its way into economic research in recent years, with early applications in the field of finance. Tetlock (2007), for example, uses a sentiment indicator obtained from news media content to predict stock price movements. His indicator builds on an "extremely rudimentary measurement rule" (Tetlock, 2007, p. 1144) which counts the occurrence of words from a predefined dictionary of 77 sentiment dimensions in news articles. Using the main principal component from the 77 sentiment dimensions, he can demonstrate that stock market prices show short-term reactions to pessimistic news reporting. Other research areas in the broader field of economics have also utilized language data in combination with simple information extraction approaches. Baker et al. (2016), for instance, use newspaper articles to measure economic policy uncertainty by means of simple keyword searches. Hoberg and Phillips (2016), in turn, construct time-varying industry classifications of companies based on heuristic vectorization techniques of the firms' 10-K business descriptions. However, the latest advances in the field of Natural Language Processing (NLP) have brought up model architectures that go beyond simple count heuristics and allow the retrieval of much more complex linguistic characteristics such as context or sentiment of word sequences. Moreover, modern NLP greatly leverages the concept of transfer learning[a] which allows researchers for the extraction of fine-grained signals from language data at low cost and processing power.

The major methodological advances in turning unstructured and often large scale data into fine-granular signals have come from the field of statistical learning - more prominently termed as machine learning in recent years. The very principle of statistical learning methods is to model a predictive, flexible functional form, $f(\mathbf{x})$, to understand unstructured and often high-dimensional data, $\mathbf{x}$, also referred as feature space (James et al., 2013). Depending on the input data, statistical learning can be broadly categorized into supervised and unsupervised learning. In supervised learning, the functional form learns the correspondence between an outcome variable, y, and the feature space. This includes regression and classification tasks. Unsupervised algorithms aim at learning how the variables in the feature set are

[a] Transfer learning in NLP aims at pretraining language models on massive text corpora and in resource-intensive processes to acquire general knowledge of the statistical properties of language. Typically, the pretrained models have learned so called distributed (vector) representations of words and sentences that can be leveraged for downstream NLP tasks like text classification. Depending on the language domain of the downstream task, the pretrained model can either be used as it is, or it can be fine-tuned based on a relatively small amount of task- and domain-specific training data to achieve even better results at low computational cost.

organized and relate to one another, e.g., through clustering.

$$\hat{f}(\mathbf{x}) = \begin{cases} \hat{y} & \text{given pairs of } (y, \mathbf{x}) \quad \text{supervised learning} \\ cluster_j & \text{given } \mathbf{x} \quad\quad\quad\quad\quad \text{unsupervised learning} \end{cases}$$ [b]

Economists put typically more structure on their models since they are interested in parameter estimation to understand the relationship between an outcome variable, y and a *specific* explanatory variable, x:

$$f(x, \mathbf{z}) = \beta x + \gamma \mathbf{z} + \epsilon \longrightarrow \hat{f}(x, \mathbf{z}) \longrightarrow \hat{\beta}$$ [c]

With its focus on function estimation as predictive tool, statistical learning has its roots in the fields of computer science, statistics and engineering (Einav & Levin, 2014) and has just begun to draw attention by economists in more recent years. In fact, renowned economists such as Susan Athey, Matthew Gentzkow, Guido Imbens, Sendhil Mullainathan or Hal Varian, among others, have made calls to encourage economic research to extend its toolkit by incorporating advanced methods from the field of statistical learning into empirical research designs (Varian, 2014; Mullainathan & Spiess, 2017; Athey & Imbens, 2019; Gentzkow et al., 2019).

Following this call, this thesis shows that statistical learning methods serve as an effective tool to tackle economic research questions that revolve around the behavior of firms and economic policymakers under dynamic market conditions. The empirical designs of the following three articles, which form the core of my dissertation, are all based on statistical learning methods. In my first article, I apply a supervised learning approach to evaluate the early policy response to the COVID-19 pandemic in Germany. Matching companies whose credit rating update is observed after the onset of the pandemic with credit rating updates prior to the crisis, sheds light on how policy measures affected closure dynamics in Germany. In a second paper, my research further shows that statistical learning allows to categorize communication patterns of companies in times of economic crises into meaningful impact indicators. These indicators entail a leading signal of later changes in the companies' credit standing. In highly dynamic times, such as at the beginning of the Corona pandemic, policymakers can benefit from such leading indicators to overcome information deficits and to design their aid measures most effectively. Moving from closure dynamics to entry dynamics, my third paper focuses on the role of market entrants in the diffusion of environmental technologies. It is shown how textual innovation data from patents and company descriptions can be leveraged to

[b] with $j \in 1, \ldots, J$ and J as number of distinct clusters. Besides clustering, unsupervised learning includes also dimensionality reduction techniques such as principal component and factor analysis.

[c] with $\mathbf{z}$ as set of control variates.

statistically learn a venture's orientation towards sustainable technology solutions. Measuring a company's technological profile is particularly challenging for the case of newly founded firms for which historical innovation data does typically not exist. Since newly registered firms are legally obliged to publish their business purpose, the proposed approach opens up a new avenue to systematically learn about market entrants' technology usage. Again, this can serve as useful policy tool to direct innovation and technological change into socially desirable pathways through targeted support of start-ups that act as accelerators of a green technology transition. Summarizing, my research is characterized by the recurrent methodology of statistical learning that finds its application in economic research questions related to

- the assessment of firm dynamics under changing market conditions
- and the role of policy under such conditions.

The remainder of this thesis proceeds as follows. In Chapter 2, I present two research papers that revolve about the company closure dynamics at the onset of the Corona pandemic in 2020 and the role of policymakers that aimed at minimizing corporate insolvencies in this unprecedented shock. Chapter 3 introduces a third article that focuses on the role of market entrants in a policy-induced, directed technical change towards a desirable long-term equilibrium of green growth. For each paper in Chapters 2 and 3, I provide a short summary outlining the main contribution and findings. In Chapter 4, I discuss and connect the research strings in the papers and draw concluding remarks. All papers are available in the Appendix.

Chapter 2

Applications of Statistical Learning to Closure Dynamics during the COVID-19 Pandemic

Much of my dissertation work took place during the Corona pandemic, which not only severely impacted the health care system, but also had far-reaching economic consequences. Understanding the economic impact of COVID-19 and its related policies and guiding policymakers in their response measures became an important task for economic research. From a policy perspective, COVID-19 brought two major challenges: (1) it forced policymakers to balance between public health and economic stability and (2) it required policymakers to react swiftly, often under immense information deficits due to the historically unprecedented circumstances. This highly dynamic setting produced numerous urgent research questions and fueled the need for policy guidance through empirical evidence. It is therefore not surprising that many renowned peer-reviewed journals in the field of economics have published special issues on the various economic impacts of COVID-19. A dedicated online journal, "COVID Economics", was even established by the Centre for Economic Policy Research (CEPR) to disseminate new findings on the economic impacts of the pandemic in a timely manner. These circumstances also provided a great opportunity to demonstrate the power of statistical learning approaches, both as an empirical strategy for evaluating policy responses to the pandemic and to assist policymakers with timely insights that traditional economic data could not provide in such a highly dynamic situation. The common theme of the two articles presented in this chapter is the application of statistical learning methods to shed light on the economic dynamics triggered by COVID-19.

In the first article in this chapter, my co-authors and I analyze the policy response to the COVID-19 pandemic that aimed at preventing a wave of business insolvencies. In the paper, we examine the extent to which the policy response induced a backlog of business insolvencies. These closure dynamics have been atypical since financial distress in times of crises usually require financially weak firms to exit the market causing insolvency numbers to rise. The paper was written as part of a research project analyzing the economic effects on German firms funded by the Federal Ministry of Economic Affairs (BMWi) (grant agreement number 15/20). In addition, the study received research funding from the European Union Horizon 2020 Research and Innovation Action for the "GrowInPro" project (grant agreement number 822781).

The second paper in this chapter is concerned with the information deficits that policymakers were confronted with in the early phase of the pandemic. It proposes a data framework that combines timely online sources with more traditional policy data, such as business surveys, in order to overcome information deficits and to guide policy decisions more effectively. In the paper, we show how online communication patterns of firms serve as leading indicators of subsequent changes in their creditworthiness. Moreover, it is demonstrated that these indicators disclose the heterogeneity of the pandemic's impact across sectors at near real-time. More traditional information sources such as business surveys could only reveal these insights with a substantial time lag. The study extends upon the BMWi project mentioned before (grant agreement number 15/20). Moreover, the project received support from the Ministry of Science, Research and the Arts of the government of Baden Wuerttemberg as part of its Science Data Center program under the grant "Business and Economic Research Data Center (BERD)".

In the following, I will introduce the papers in more detail. The published full-text articles can be found in the Appendix.

2.1 Small Firms and the COVID-19 Insolvency Gap

Preface

In the paper "Small Firms and the COVID-19 Insolvency Gap", joint with Georg Licht and Simona Murmann and published in *Small Business Economics* (2022; 58:887–917), we analyze whether the policy aid measures in light of the Corona pandemic triggered a backlog in corporate insolvencies possibly impeding the Schumpeterian cleansing effect that is typically observed in times of economic shocks. The publication can be found in Appendix A. My own contribution to the publication is 60%. The source code is open access and can be found on GitHub.

Idea and Motivation

With the outbreak of the Corona pandemic and rapidly rising case numbers in early 2020, policymakers were forced to implement unprecedented containment measures including the temporary shutdown of a wide range of economic activities. The resulting slump in sales and revenues, coupled with unchanged fixed cost obligations, posed an existential threat to many companies, especially to smaller ones with only limited financial reserves. At the time, there was intense public speculation about a possible "wave of bankruptcies" and policymakers were challenged to prevent precisely such a scenario at all costs (see, e.g., The Economist (2020)). For this reason, the Federal Government launched a series of mainly indiscriminate aid measures deemed as the "largest assistance package in the history of the Federal Republic of Germany" (Federal Ministry of Finance, 2020, p. 3). As a result, insolvency figures fell to their lowest level since the introduction of the German insolvency law - despite the deep economic recession. These atypical dynamics of insolvency filings gave rise to the assumption that the fiscal policy response led to an impediment of the cleansing mechanism that is typically observed during times of economic shocks. Following the pioneering thoughts of Joseph Schumpeter, economic crises typically force unviable firms to exit the market which leads to a reallocation of their resources to more efficient companies (Schumpeter, 1942). The objective of this paper is to empirically assess whether the economic policy assistance in the context of the COVID-19 crisis held back insolvency filings and to analyze whether this kept firms alive that had already been under financial distress before the pandemic.

Data and Methodology

We use company-specific credit ratings in conjunction with information on insolvency filings to estimate the number of insolvencies that would have been expected in the absence of the policy aid measures. Credit ratings reflect a firm's creditworthiness and are commonly used to predict corporate bankruptcy both in practice and in research (Altman, 1968, 2013). Observing rating updates before and after the pandemic for the near universe of economically active companies in Germany allows for the creation of a crisis period, spanning the first months after the lockdown measures in 2020, and a control period, spanning more than two years prior to the outbreak. In this way, it is possible to compare rating updates in times of strong political support with rating changes of similar companies, but in a situation where politics did not intervene to save companies from bankruptcy. See Figure 2.1 for an illustration. We then track the company's survival status in the months after it received its rating update for a large sample of credit rating updates observed in both periods.

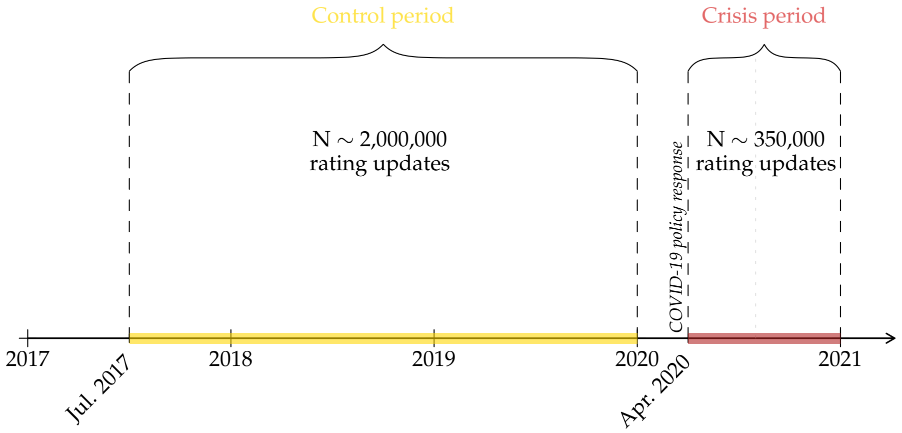


FIGURE 2.1: Sample of credit rating updates split by crisis and control period.

The core challenge is to statically learn "similarity" between rating updates in the control period and the crisis period. For this purpose, we match the k nearest neighbors from the control period to each firm credit rating update observed during the crisis period as illustrated in Figure 2.2. For determining the nearest neighbors, we only consider firms from the same industry sector and within the same size class. Within each sector-size stratum, Mahalanobis Distance (MD) on observable characteristics, $\mathbf{x}_i$, such as the change in rating points, the rating level prior to the update or the company's age, between each possible pair of control and crisis unit is calculated to find the k nearest neighbors.

Observing the insolvency state of the focal crisis observation with the insolvency states of its nearest neighbors from the the pre-crisis period allows for the derivation of an insolvency probability given the focal firm's rating update.

$$\hat{f}(\mathbf{x}_i) = Pr\left(f_{i,t+z} = 1 \mid \mathbf{x}_i\right) = \frac{1}{k} \sum_{j \in N_k(\mathbf{x}_i)} \mathbf{1}(f_{j,t+z} = 1)$$ [a]

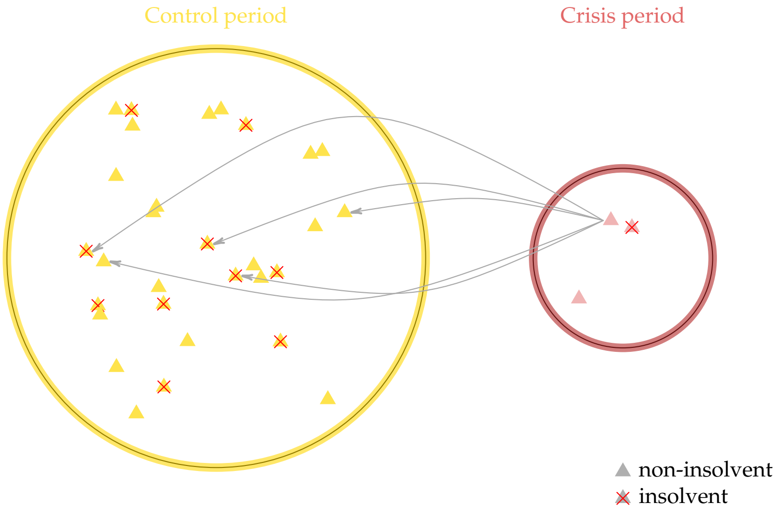


FIGURE 2.2: Illustration of matching approach via *k*NN. Here *k*=5.

Translating this idea on an aggregated view for each sector-size stratum, s, allows for the comparison of actually observed insolvency rates, IR_s^{actual}, in the first months after the outbreak with counterfactual insolvency rates that would have been expected without the policy intervention, $IR_s^{counterfactual}$. The deviation between actual and counterfactual insolvency rates provides an empirical estimate of the backlog of corporate insolvencies, the so called insolvency gap, IG, for each sector-size stratum.

$$\begin{aligned} IR_s^{actual} &= \frac{N_s^{insolvent}}{N_s} \\ IR_s^{counterfactual} &= \frac{\sum_{j=1}^{\tilde{N}_s} w_{j,s}\,\mathbf{1}(f_{j,t+z}=1)}{\sum_{j=1}^{\tilde{N}_s} w_{j,s}} = \frac{1}{N_s}\sum_{i=1}^{N_s} Pr\left(f_{i,t+z}=1 \mid \mathbf{x}_i\right)^{b} \\ IG_s &= IR_s^{counterfactual} - IR_s^{actual} \end{aligned}$$

RESULTS

Our estimates suggest that the aid measures led to a substantial insolvency gap among very small firms with 10 employees or less as displayed in Table 2.1. With the

[a] $f_{i,t+z}$ reflects the survival status of firm i z-months after its rating update. $N_k(\mathbf{x}_i)$ determines the k closest points in the neighborhood of $\mathbf{x}_i$ learned via MD:

$$MD_{ij} = \begin{cases} (\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j) & \text{if } \quad |\Delta r_{it} - \Delta r_{jt}| \leq c \\ \infty & \text{if } \quad |\Delta r_{it} - \Delta r_{jt}| > c \end{cases}$$

with Σ as the variance covariance matrix of $\mathbf{x}$ in the pooled sample of in-crisis and all pre-crisis observations, Δr_{it} as the credit rating update of firm i at time t and c as strict caliper ensuring that only companies which experienced almost the same rating update will be matched.

[b] with N_s as number of observed companies in the crisis period in s, $\tilde{N}_s$ as number of matched pre-crisis observations in s and $w_{j,s}$ as matching weight on j in s.

decline in trading activity and the lack of business revenue, many companies have had to rely on their cash reserves to meet their unchanged fixed cost obligations. Because smaller and entrepreneurial firms are characterized by a strong reliance on internally generated funds to capitalize their businesses, both their cash reserves and collateral for borrowing are usually limited. In times of financial distress, such as at the beginning of the COVID-19 crisis, this made smaller businesses particularly vulnerable to financial insolvency as recorded in their ratings' deterioration. Thus, micro firms were the ones which benefited most from the generous policy aid measures that kept them alive despite their financial distress.

Sector affiliation	**Size of company**			
	Micro $\widehat{IG}_s$	Small $\widehat{IG}_s$	Medium $\widehat{IG}_s$	Large $\widehat{IG}_s$
Manufacturing	1.033***	0.019	−0.041	−0.350
Business-related services	0.704***	−0.007	−0.053	0.000
Food production	0.274	0.242	−0.188	−1.050**
Others	0.370***	−0.018	0.000	0.000
Manufacturing of data processing equip.	0.442*	−0.090	0.000	−1.220*
Mechanical engineering	0.033	0.177	−0.246***	0.000
Accommodation & catering	1.147***	0.053	0.276	0.000
Creative industry & entertainment	0.123	0.173	0.000	0.000
Health & social services	0.370***	0.053	−0.115	0.040
Insurance & banking	0.370***	0.000	0.000	0.000
Logistics & transport	0.704***	0.021	0.298	0.000
Chemicals & pharmaceuticals	0.328*	0.030	0.000	0.000
Wholesale & retail trade	1.074***	0.040	0.007	−0.060

TABLE 2.1: Insolvency gap estimates presented in percentage points (pp). Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$ based on χ^2-Test for equality in the insolvency proportions using Rao-Scott corrections to account for matching weights. Size classes according to number of employees: Micro: ≤ 10, Small: 11-49, Medium: 50-249, Large: ≥ 250. From Dörr et al. (2022b).

Distinguishing between firms with a decent rating and those firms with a below average rating prior to the pandemic, we further find that the gap is particularly prevalent among firms which had a comparatively bad financial standing already before the crisis. This can be seen in Figure 2.3. It becomes obvious that the insolvency gap is strongly driven by micro companies (10 employees and less) with a weak financial standing already before the onset of the pandemic. This result suggests that financially weak firms remained in the market, possibly absorbing the relief measures as windfall gains. In light of these findings, we argue that well-dosed and targeted liquidity injections would have been needed as a more selective measure to support firms in the early stages of the pandemic, rather than choosing the "bazooka" (Federal Ministry of Finance, 2020, para. 1) that provided relief on a lump-sum basis without taking into consideration the firms' pre-crisis conditions. At the same time, we acknowledge that the urgency of financial aid measures gave politics little time to conduct extensive credit assessments. Clearly, this has shown that timely information on how different firms are impacted by an economic shock is necessary to allow for an differentiation of public aid measures.[c]

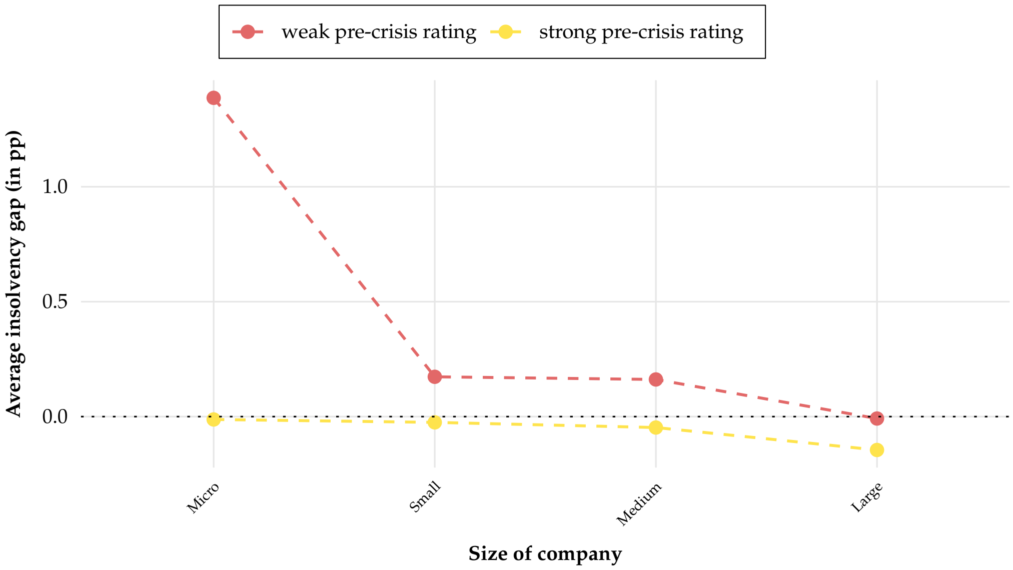


FIGURE 2.3: Average insolvency gap estimates by size class and financial standing predating the crisis. From Dörr et al. (2022b).

Conclusion and Outlook

We analyzed the effect of the largely indiscriminate aid measures of the German Federal Government provided to companies in the first months of the COVID-19 pandemic. While the government's goal was to prevent a wave of corporate bankruptcies due to the necessary shutdown measures, there was also the risk that unviable firms could absorb public liquidity aid to remain in the market. In times of crises, insolvencies usually allow for a reallocation of employees and capital to more efficient firms. Hampering this cleansing mechanism could have long-term effects on aggregate productivity. Now, after more than two years into the pandemic, it can be said that the government has successfully prevented a wave of insolvencies even after the aid measures had expired, since the number of insolvencies has remained

[c] The article presented in Chapter 2.2 is devoted to this issue.

at a low level to date. Future research will show how aggregate productivity has been impacted by the pandemic.

2.2 An Integrated Data Framework for Policy Guidance during the Coronavirus Pandemic: Towards Real-Time Decision Support for Economic Policymakers

Preface

In the article "An Integrated Data Framework for Policy Guidance during the Coronavirus Pandemic: Towards Real-Time Decision Support for Economic Policymakers" that is co-authored with Jan Kinne, David Lenz, Georg Licht and Peter Winker and published in *PLoS ONE* (2022; 17(2):e0263898), we propose a data framework for guiding policy decisions. Especially in highly dynamic situations, such as in the first months of the Corona pandemic, the framework provides leading indicators on the economic impact of the pandemic at firm level. The publication can be found in Appendix B. I contributed with 60% to the publication. The interested reader can find the paper's source code on GitHub.

Idea and Motivation

The course of events following an economic shock is usually highly uncertain. At the beginning of the Corona pandemic, when swift policy response was crucial to prevent business failures of otherwise healthy companies, this uncertainty posed a great challenge for politics. Economic policymakers were on the one hand forced to act quickly to cushion the economic impact of the pandemic and its shutdown measures. On the other hand, they faced major information deficits regarding the dynamics and impact heterogeneity of the shock. As a consequence, many governments saw no other chance than to intervene in the economy on an enormous scale, mostly in an untargeted manner and at high fiscal cost. The German government, for example, fired the much-cited "bazooka" (Federal Ministry of Finance, 2020, para. 1) of aid measures in the wake of the pandemic and provided large lump-sum liquidity support.

In this regard, the pandemic has revealed that the "gap between official data and what is happening in the real economy can still be glaring" which is why "many policymakers have operated in a fog" (The Economist, 2021, para. 11-12). Traditionally, policymakers rely on official statistics, business surveys or ex-post evaluations of previous policy measures to decide on effective spending of public resources. However, in the highly dynamic situation that followed the virus outbreak, none of these sources of information seemed helpful. This is because they require long lead times for preparation, execution and analysis like in the case of surveys. Also learning from past experiences was not possible due to the unprecedented nature of the economic dynamics caused by the pandemic. Other data sources, such as balance sheet or credit rating data, pose another source of important indicators. But they, too, have the disadvantage that they only reflect the effects of a crisis when the economic shock has already reached its full impact and has materialized in payment discipline and other performance indicators. Tackling this lack of real-time data in times of economic shocks, this paper presents a data framework that combines traditional policy data with webdata that is being disseminated continuously along the dynamics of the shock. More precisely, the article shows how firm level communication patterns about the pandemic retrieved from corporate websites allowed for disclosure of the heterogeneity of the pandemic's impact and served as a leading indicator for changes in the creditworthiness of firms at near real-time.

Data and Methodology

Three days after the announcement of the nationwide shutdown in Germany on March 16, 2020, we began accessing the corporate websites of nearly 1.2 million German companies twice a week, searching for references to the pandemic. This approach revealed that German companies used their websites intensively to report about the pandemic as displayed in Figure 2.4.

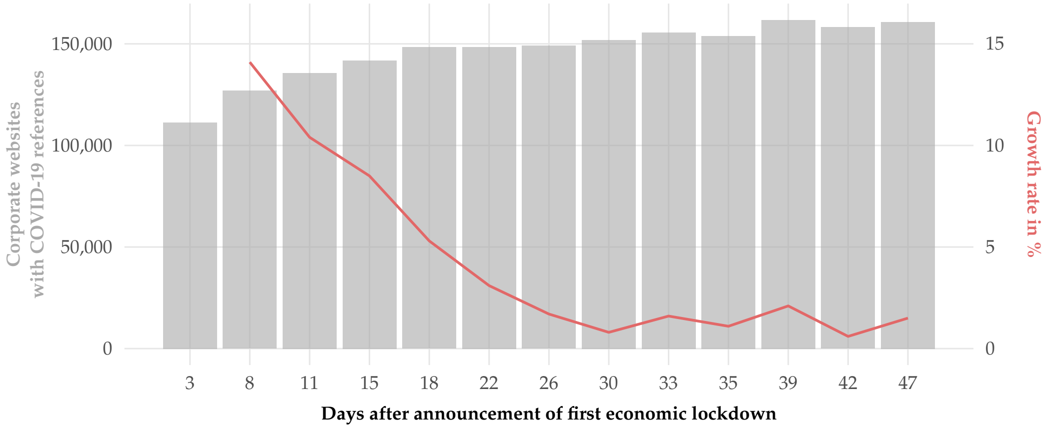


FIGURE 2.4: Number of firms which reported about COVID-19 on their corporate websites (left), percentage increase of this number over time (right). From Dörr et al. (2022a).

The context in which companies communicated about the pandemic varied widely. Some reported about problems (closures, cancellation of events, short-time work, etc.), others announced adaptive measures (adjusted opening hours, delivery service, home-office regulations, etc.). In order to retrieve meaningful signals from these communication patterns, we have introduced five distinguishable impact classes: (1) problem reports, (2) confidence in dealing with the crisis, (3) adaption to the pandemic, (4) non-business related information about the pandemic and (5) others. Given textual references of more than 150,000 companies, the statistical learning task was to derive a functional form that allowed for a classification of all of the found references into one of the five impact classes. For this text classification task, we have leveraged the power of transfer learning which has become the quasi-standard for many NLP tasks (Howard & Ruder, 2018; Ruder et al., 2019). Based on a manually labeled training data set, we have fine-tuned XLM-RoBERTa (Cross-lingual Language Model - Robustly optimized BERT approach) (Conneau et al., 2019), a pretrained language model that encodes sentences into distributed vector representations. Fine-tuning the model allows for the transfer of the model's general knowledge of linguistic patterns and regularities learned during pretraining to the specific task of classifying the textual references, $\mathbf{x_i}$, into one of the five context classes, y_i.

$$\hat{f}\,(\mathbf{x}_i) = \textit{majority vote}\,\{\text{XLM-RoBERTa}_e(\mathbf{x_i})\}_1^E = \hat{y}_i \in \{(1),(2),(3),(4),(5)\}$$ [d]

Enriching the company-specific impact classes with a traditional business survey on the economic effects of COVID-19 and firm-specific credit rating data, as shown in Figure 2.5, complements our proposed data framework for guiding policy decisions. Most importantly, this framework focuses on bridging information gaps by highlighting heterogeneity across all phases of an unanticipated economic shock. Continuously updating policymakers' information set once new information sources become available enables more surgical policy actions than implemented in the early stage of the pandemic. Secondly, by combining these data sources, we can show that online sources from the internet, disseminated in real-time, can be transformed into leading indicators that reflect how a shock materializes in the economy. This is demonstrated in the following.

[d] Note that in the paper, $\hat{f}$ is an ensemble of E different versions of XLM-RoBERTa fine-tuned on distinct subsamples of the training data. The final classification is based on a majority decision rule.

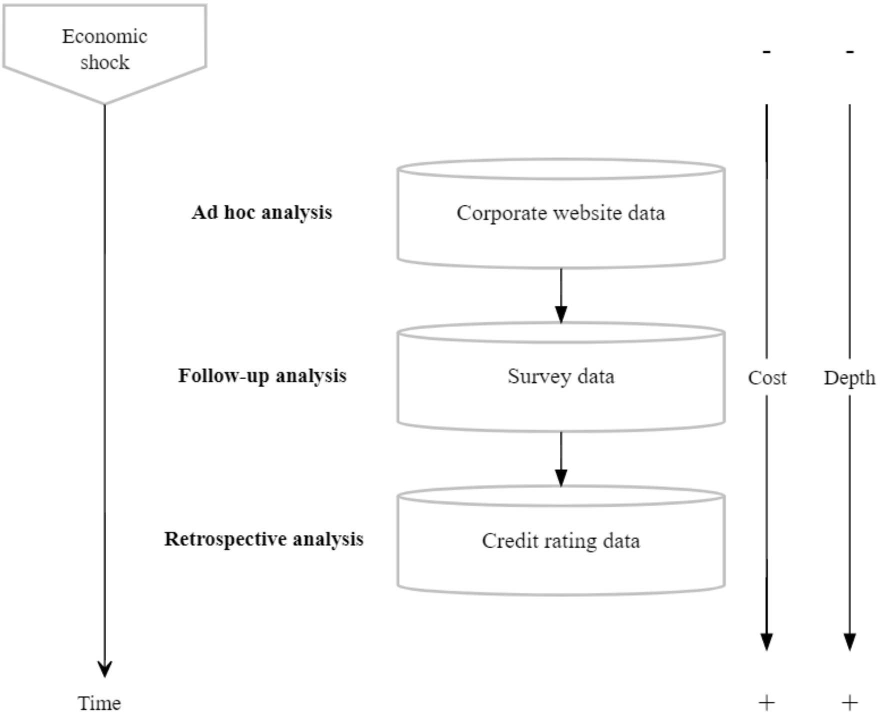


FIGURE 2.5: Data framework for tracking effects of economic shocks on businesses by combining corporate website, business survey and credit rating data. From Dörr et al. (2022a).

RESULTS

The results in the article show that the classified text references can equally capture the heterogeneity of the shock as revealed by conventional business surveys. The key difference, however, is that survey results become only available several weeks after the shock, whereas webdata can be collected in real-time. For this purpose, we regress company characteristics, such as age, size and sector affiliation[e] on a dichotomous variable indicating whether the company has been negatively exposed to the pandemic. In the webdata sample, these are companies whose text references were classified as $\hat{y}_i = (1)$ problem; in the survey, these comprise companies that responded that they faced one of the following problems: decline in demand, temporary closure, supply chain disruption, staff shortages, logistical sales problems, or liquidity constraints. Figure 2.6 shows that the effect estimates based on the webdata sample closely resemble the effect estimates from the business survey.

Table 2.2 shows that the impact classes retrieved in near real-time from the sample of company websites also entail a leading indication on how credit ratings have

[e] Company characteristics are available via the Mannheim Enterprise Panel (MUP).

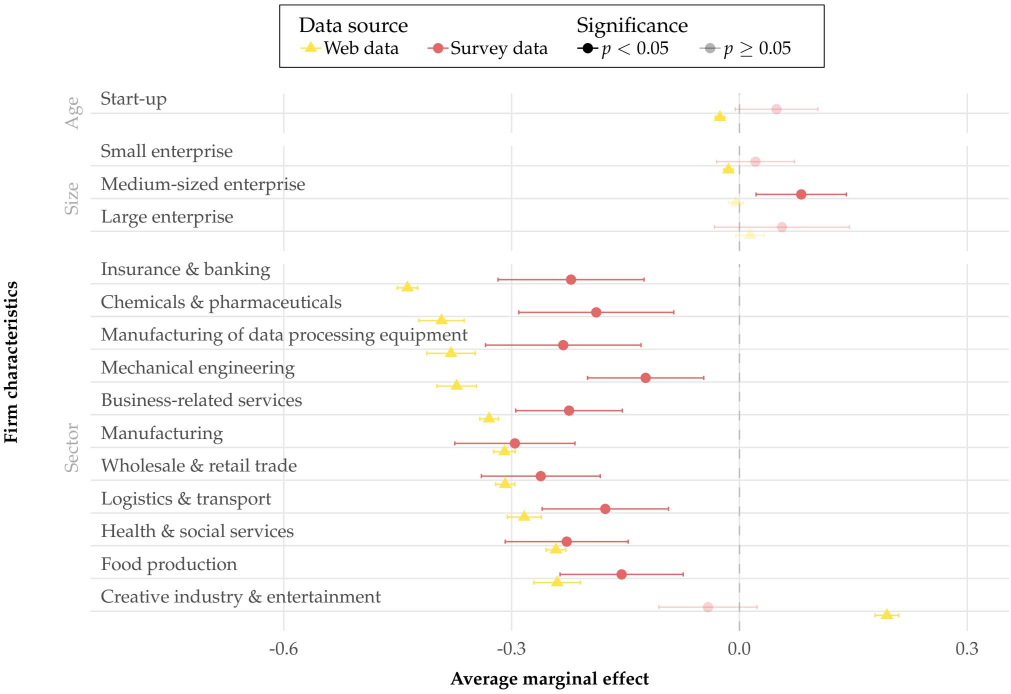


FIGURE 2.6: Comparison of webdata and survey data effect estimates and corresponding 95%-confidence intervals. Dependent variable reflects whether company signaled that it is negatively exposed to the crisis, either by having communicated this through its website (yellow) or by a respective response in the survey (red). Incumbent firms (10 years and older) serve as baseline age group, micro-enterprises (number of employees ≤ 10) as baseline size group, accommodation and catering serves as baseline sector among the sector controls. From Dörr et al. (2022a).

changed in response to the pandemic. Results are based on regressing firm-specific credit rating updates after June 01, 2020, $\Delta r_{\bar{t}+z}$, on impact indicators generated from communication data retrieved before June 01, 2020.[f] After controlling for age, size and sector fixed effects, in model specification (4), companies that reported facing pandemic-related problems experienced on average a statistically significant deterioration in their credit rating later on in the crisis. This suggests that early communication patterns from online sources can serve as leading indicators of how an economic shock will play out at firm level, as in the form of changes in creditworthiness, for instance.

[f] Note that the context classes have been extracted from corporate websites between March 2020 and May 2020, i.e., before June 01, 2020. We express this time period with the index $\bar{t}$. $\bar{t} + z$ reflects the date of rating update with z as number of days after $\bar{t}$.

	$\Delta r_{\bar{t}+z}$			
	(1)	(2)	(3)	(4)
Problem$_{\bar{t}}$	$+1.66^{***}$ (0.18)	$+1.68^{***}$ (0.18)	$+1.62^{***}$ (0.19)	$+0.42^{**}$ (0.19)
Confidence$_{\bar{t}}$	-1.70^{***} (0.42)	-1.69^{***} (0.42)	-1.73^{***} (0.43)	-0.69 (0.43)
Adaption$_{\bar{t}}$	-0.46^{***} (0.08)	-0.47^{***} (0.08)	-0.33^{***} (0.08)	-0.13 (0.08)
Information$_{\bar{t}}$	-0.24^{***} (0.04)	-0.24^{***} (0.04)	-0.23^{***} (0.04)	-0.17^{***} (0.04)
Other$_{\bar{t}}$	-0.42^{***} (0.12)	-0.42^{***} (0.12)	-0.10 (0.12)	-0.08 (0.12)
Rating level predating pandemic	-0.09^{***} (< 0.01)	-0.10^{***} (< 0.01)	-0.11^{***} (< 0.01)	-0.13^{***} (< 0.01)
Age controls	No	Yes	Yes	Yes
Size controls	No	No	Yes	Yes
Sector controls	No	No	No	Yes
N	61,228	61,138	57,343	57,343

TABLE 2.2: Text references on corporate websites as early indicators for changes in firm credit ratings. Effect estimates presented in rating index points (rating index ranges from 100 to 500 with a higher index signaling a worse financial standing). Positive sign estimates reflect rating downgrades, negative ones signal upgrades. White robust standard errors are reported in parentheses. Change in observation numbers due to missing information about firm characteristics. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$. From Dörr et al. (2022a).

CONCLUSION AND OUTLOOK

This paper has shown that online sources, such as corporate website data, serve as valuable information source that sheds light on the impact of economic shocks at *firm level* and at *near real-time*. High granularity and timeliness of this form of information, often channeled in an unstructured format through the internet, make it a valuable source to complement traditional economic data for guiding policy decisions. This becomes particularly relevant when more traditional data sources lack timeliness but political decision-makers need to respond swiftly like in the first months after the Corona outbreak. The article has shown that it is possible to create leading impact indicators from such unstructured forms of online data using methods from the field of statistical learning. This is a first step towards real-time decision support for economic policymakers bridging information deficits in dynamic situations. As important guide to public-sector decision-making, economists should continue investigating how timelier and more granular information sources can produce reliable insights when consulting political decision-makers.

Chapter 3

Applications of Statistical Learning to Entry Dynamics under Directed Technical Change

Besides the unforeseen dynamics of the Corona crisis, climate change has been and continues to be an additional source of market reordering that poses immense challenges for policymakers and society. With its consequences becoming increasingly apparent, policymakers are seeking to regulate companies and industries with the goal of reducing their environmental footprint. Carbon pricing, for example, is intended not only to help reduce industrial emissions, but also to induce the development of new, climate-friendly technologies (Calel & Dechezleprêtre, 2016). This is one example that demonstrates how politics aims at tackling climate change by redirecting technology regimes into environmentally-friendly pathways. Directed technological change in environmental policy is based on the belief that we can address anthropogenic climate change through innovation and technology. Delivering these technological innovations is a great opportunity for businesses and promoting their development is a central task of 21^{st} century policy-making.

For economic research to evaluate how successful instruments of directed technological change are, it is necessary to be able to measure technology usage and technology change within firms. This allows for an understanding which types of firms drive technological change and respond to policy measures that aim at inducing such change. From a theoretical standpoint, directed technical change in environmental policy not only induces existing firms to eco-innovate, but also creates incentives for entrepreneurial activity to take advantage of changing market demands, ultimately affecting the dynamics of market entry. With this in mind, the article presented in this chapter focuses on the role of new entrants, to whom theory assigns a special role in the diffusion of climate change mitigation technologies (see, e.g., Hockerts and Wüstenhagen (2010)). Independent from past investment decisions, start-up firms are seen as important accelerators of environmentally sound market solutions. They do not face technological path dependencies that pose barriers for incumbent firms in radical environmental innovations. To enable policymakers to assess the technological orientation of start-up companies, the paper shows how textual information about a company's business model can be used to statistically learn its technology profile. This helps policymakers to narrow down market entry dynamics in the clean technology sector. Ultimately, the article empirically examines the innovation characteristics of clean technology-oriented start-ups to judge whether they are in line with theory on technological path creation through start-up firms. In the following, I will present the paper in more detail. The full-text paper can also be found in the Appendix.

3.1 Mapping Technologies to Business Models: An Application to Clean Technologies and Entrepreneurship

Preface

In the article "Mapping Technologies to Business Models: An Application to Clean Technologies and Entrepreneurship", I propose a novel approach to asses start-ups' technological orientation based on their business descriptions that they are typically obliged to report upon business registration. The working paper is under review for the 26th *International Conference on Science, Technology and Innovation Indicators* (STI 2022). The full-text paper can be found in Appendix C. The paper's source code is available on GitHub.

Idea and Motivation

While much is known about environmental innovation and clean technology diffusion by incumbents, there is little empirical research that allows for a better understanding of start-ups' role in the technological transition to a low-carbon economy. This is mainly because there is no historical data on R&D and patenting activities of new companies, which is required for determining their innovation potential and technological orientation. From a theoretical perspective, however, start-up companies are attributed a special role in creating new technological pathways that accelerate the adoption and diffusion of sustainability innovations (see, e.g., Hockerts and Wüstenhagen (2010)). Unconstrained by previous technological investments, entrants can introduce more radical environmental innovations than incumbent firms that are often locked into path dependencies by their previous technology choices. Yet, the importance of start-ups as technological path creators towards higher sustainability standards has been derived from theoretical considerations and are conceptual or anecdotal in nature (Olteanu & Fichter, 2022). The lack of an empirical basis for identifying clean technology-focused entrepreneurship makes it difficult to back their special role as transition enablers with empirical evidence. From a policy perspective, it also hampers policymakers to effectively promote the type of entrepreneurship that spurs innovation towards technologically desirable pathways.

To enable empirical identification of clean technology-focused start-ups, this paper proposes an approach that allows for a systematic mapping of start-up business models on a set of well-defined clean technologies based on textual innovation data. The framework demonstrates that, despite the sparse information base of newly formed ventures' innovation capacity, it is still possible to learn something about the relevance of specific technologies for the firm. To this end, the approach uses company descriptions that start-ups must provide when registering their business.

It is shown that from this textual source of business information it is possible to measure a firm's technological orientation. An application of the framework to a survey of German start-up firms suggests that cleantech-oriented market entrants exhibit innovation characteristics consistent with theory on technological path creation. The framework not only has the potential to address new research questions in the area of environmental innovation, but also provides policymakers with a tool to tailor subsidies, tax incentives, and other start-up support programs in favor of ventures with a high potential to accelerate green technological change.

Data and Methodology

The proposed approach relies on two sources of textual innovation data: business descriptions of newly registered ventures and patent texts. Methodologically, measuring a venture's orientation towards a particular set of technologies is derived following two steps. In a first step, technical terms that refer to narrowly defined technologies are filtered from a large corpus of patent texts. Besides the textual content of the patents, this step also leverages information on the patents' technology classes that patent examiners assign as part of the patent granting procedure. Based on this labeled corpus of patent texts, I use Labeled Latent Dirichlet Allocation (L-LDA) (Ramage et al., 2009), a supervised topic model, to learn technology-specific terms from the corpus. More precisely, L-LDA models the patent corpus by a joint probability distribution over both the observed terms, w, in each patent document but also hidden variables such as the technologies, $\delta_{1:T}$, expressed as distribution over the corpus' vocabulary.[a]

$$p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P}) = \prod_{t=1}^{T} p(\delta_t) \prod_{p=1}^{P} p(\lambda_p) \left(\prod_{n=1}^{N_p} p(z_{p,n}|\lambda_p) p(w_{p,n}|\delta_{1:T}, z_{p,n}) \right)$$ [b]

To derive semantic representations of technologies the statistical learning goal is to derive their probability distributions, $p(\delta_t)$, over the corpus' vocabulary, $\mathbf{x}$. In other words, it is the task of inferring the marginal distributions, $p(\delta_t)$, from the joint probability distribution for each of the technology classes t. For this purpose, δ_t is modeled as V-dimensional Dirichlet random variable (Ramage et al., 2009) with V as the number of distinct vocabulary in the patent corpus. Retaining only the top-Q terms with highest probability translates into a sequence of technical terms[c] that

[a] Additional hidden variables comprise the distribution of technology fields over patent abstracts, λ_p, and the technology assignment for the n^{th} word in patent p, $z_{p,n}$ (Blei, 2012).

[b] T: number of technology classes in corpus, P: number of patents in corpus, N_p number of words in patent p.

[c] Note that the sequence is ordered by descending probability with which the technical terms describe the corresponding technology. Keeping only the Q terms with highest probability effectively filters the technical terms that are most relevant in the describing the focal technology. In the paper,

describe the focal technology:

$$\hat{f}(\mathbf{x}) = \langle w_q : q \leq Q \rangle_t \quad \forall\, t \in \; \{1, \ldots, T\}.^{\text{d}}$$

In a second step, the word sequences are transferred into a distributed representation using Sentence-BERT (SBERT) (Reimers & Gurevych, 2019), a pretrained text embedding model. Similar to the second paper, I transfer the linguistic knowledge that the model acquired during pretraining to map a system of distinct technologies in vector space. Converting the business descriptions of new ventures using the same embedding model into distributed representations, allows for a shift of companies into the technology space. I then propose cosine similarity as measure of technological proximity of a company's embedding, Y_i, to any of the technology embeddings, X_t, within that space.

$$\textsc{TechProx}_{t,i} := sim(X_t, Y_i) = cos(\theta_{t,i}) = max\left(0, \frac{\bar{X}_t \bar{Y}_i}{||\ \bar{X}_t\ ||||\ \bar{Y}_i\ ||}\right) \in [0, 1]$$

With this metric, it is then possible to measure how close the business model of any company is to any of the technologies that have been modeled from the patent corpus. The overall approach is illustrated in Figure 3.1.

the exact value of Q is treated as hyperparameter and is learned empirically by means of a technology-labeled dataset of company descriptions.

[d] $\langle w_q : q \leq Q \rangle_t$ is derived from $p(\delta_t)$ by sorting the terms in descending probability order and retaining only the Q terms with highest probability. For example, the word sequence for Carbon Capture and Storage (CCS) technologies derived in the paper looks as follows: $\underset{(1 \times Q)}{\delta_{CCS}} =$ $\langle$gas, absorption, dioxide, carbon, $\ldots$, scrub, separation, desorption$\rangle$.

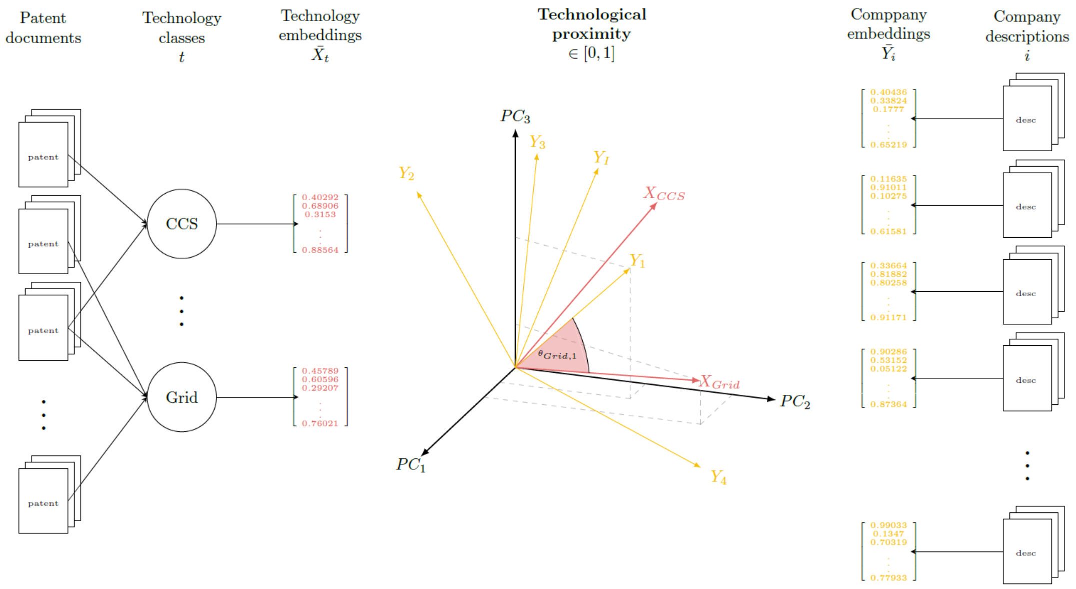


FIGURE 3.1: Illustration of framework to map technologies to company descriptions based on textual innovation data. Using technology-labeled patent texts, a vector system of technologies can be derived (red vector representations). Placing vectorized representations of company descriptions into this system (yellow vectors) it is possible to determine how closely a company's business model is located to any of the technology within the system. For this purpose, cosine similarity as angle between company embeddings and technology embeddings is proposed. For ease of illustration, embeddings are displayed on their three main components. From Dörr (2022).

RESULTS

To demonstrate that the framework allows for the detection of clean technology-oriented firms, I model technology embeddings for a well-defined set of green technologies, including CCS technologies, renewables and wastewater treatment technologies among others, from a large corpus of patent abstracts. Moreover, I leverage a dataset of company descriptions that are labeled as cleantech-oriented and non-cleantech-oriented. In this way, it is possible to assess how well TECHPROX$_{t,i}$ serves to differentiate firms whose business models are closely oriented towards green technologies from firms whose business models are unrelated to the cleantech sector. The performance metrics, as displayed in Table 3.1, show that the proposed measure of technological orientation serves well to identify cleantech firms based on their business descriptions.

Label	**Precision**	**Recall**	**F1-Score**	**Support**
Cleantech	0.87	0.86	0.86	284
Non-cleantech	0.83	0.84	0.83	233
			0.85	517

TABLE 3.1: Performance of TECHPROX in distinguishing cleantech from non-cleantech firms. Performance measured on random test set with optimal value of $Q = 15$ of most important terms used to model technology descriptions. If $\text{TECHPROX}_{t,i}$ exceeds 0.27, company i is considered as oriented towards technology t. This optimal threshold as well as the optimal value for Q have been determined on an independent validation set by tuning F1-Score. Results suggest that if the framework identifies a firm as cleantech, this is correct in 87% of cases. Overall, the framework successfully retrieves 86% of cleantech firms in the test set. From Dörr (2022).

With the main purpose of the indicator to be applicable to start-ups that are legally required to report a business purpose upon registration, I apply the framework to a sample of company descriptions of German market entrants. The sample stems from the IAB/ZEW Start-up Panel, a business survey whose wave in 2018 included questions on the surveyed firms' environmental innovation activity ($EInno$) and on the environmental impact of their products and services. In this way, it becomes possible to empirically analyze the characteristics of clean technology-oriented ventures. Table 3.2 shows that start-ups whose business model focus on clean technologies, as determined via TECHPROX, are characterized by a significantly higher propensity to introduce environmental innovations. This result is robust against the inclusion of a wide set of innovation-, performance- and product controls. Moreover, in the full-text paper it is also shown that products and services of cleantech-motivated business creations have a significant positive impact on their customers' emissions, resource consumption and recyclability. Overall, this suggests that cleantech start-ups act as path creators of green technical change: both by virtue of their existing products and services and through a high propensity to introduce additional environmental innovations.

	EInno			
	(1)	(2)	(3)	(4)
Cleantech	0.068***	0.064**	0.060**	0.078***
log(Size)	0.013***	0.012***	0.017***	0.015***
Age	0.003	0.000	0.001	0.001
Subsidy	0.067***	0.074***	0.081***	0.089***
R&D	0.081***	0.082***	0.108***	0.114***
R&D intensity	−0.055	−0.016	−0.020	−0.039
Returns		0.126***	0.111**	0.103**
Break even		0.078***	0.065***	0.071***
Team size			−0.019*	−0.023*
University			−0.122***	−0.115***
Sector controls	Yes	Yes	Yes	Yes
Product type controls	No	No	No	Yes
N	3,269	3,192	3,192	2,774
Pseudo R^2	0.037	0.043	0.054	0.061

TABLE 3.2: Clean technology-oriented entrants and their proclivity to introduce environmental innovations. Dependent variable, *EInno*, reflects whether entrant introduced an environmental innovation (reduction of energy usage, emissions or water, improved recyclability or durability of products) since its registration. Coefficient estimates reported as average marginal effects reflecting the percentage point change in the probability to introduce an environmental innovation if the explanatory variable increases by one unit. Change in observation numbers due to item non-response. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$. From Dörr (2022).

Conclusion and Outlook

In this article, I have shown that it is possible to measure technological orientation at firm level even when innovation-related data is sparse, as in the case of firms that are new to the market. Providing a structured approach that relies on advances in the field of NLP and leveraging technology-related information encrypted in companies' business descriptions, opens new gateways to better understand the role of market entrants in the technological transition towards green growth. This paper has shown that cleantech start-ups embody characteristics that favor technological path creation. Under directed technological change, the proposed measure of technological orientation may also prove as a useful policy tool to effectively promote the type of entrepreneurship that accelerates innovation towards technologically desirable pathways.

Chapter 4

Discussion

This dissertation has focused on the intersection between methods from the broad field of statistical learning and their application in empirical economic research. My research has shown that statistical learning provides powerful data modeling and data pre-processing methods that enable us to tackle economically motivated research questions that would be difficult to answer without statistical learning. At their core, statistical learning methods have in common that they optimize the parameters of a predictive function, also known as weights, through numerical optimization algorithms that minimize a predefined loss function.[a] The power of statistical learning lies in the models' ability to improve their predictive performance when trained with larger amounts of data. In the age of digitization and its ever increasing size of records on economic trans-, inter- and reactions, this advantage becomes even more obvious both in theory and in practice which explains the enormous recent methodological advances that transformed statistical learning methods to systems of artificial intelligence.

The goal of my thesis was to show that empirical economic research, too, can benefit from these advances. Whether it is determining similar pairs of companies from a large sample of firm characteristics, evaluating policy interventions during COVID-19 or transforming unstructured data like company descriptions and communication patterns into meaningful indicators to characterize technology-oriented start-ups or forecast credit rating changes. All of the three papers presented in the thesis share a connection through the application of statistical function estimation as predictive tool to understand firm dynamics under changing market conditions.

Several specific issues for future research arise from my work, which are discussed in detail in the respective articles. At this point, I would like to address two broader issues for future research that are related to my dissertation and seem worthwhile mentioning to me.

[a] An exception builds the *k*NN algorithm applied in the first paper. Instead of minimizing a loss function, the algorithm calculates the distance for all pairs of observations and selects the *k* closest observations for the underlying prediction task. Therefore, it is also referred to as distance-based statistical learning approach.

1. Clearly, parameter estimates in my thesis do not allow for causal interpretation but are correlative/predictive in nature.[b] At its core, prediction is where statistical learning shines. It offers tools for predictive function estimation to learn significant structure from typically high-dimensional or unstructured data. However, many economic research questions are concerned with the causal relationship between, e.g., a policy measure and a firm outcome. Econometrics has brought up different identification strategies that allow for causal interpretations in observational studies (Varian, 2014). Yet a recent line of research has begun to combine causal inference with statistical learning models. This is where I see great potential for future empirical research that is concerned with causal questions. For example, Chernozhukov et al. (2018) use statistical learning to adjust for differences between treated and control units in a setting where researchers are confronted with a large set of potential confounding variables. Going beyond the estimation of an *average* treatment effect, Wager and Athey (2018), leverage statistical learning for identifying and estimating *individualized* treatment effects. Their approach allows for the interpretation of effect heterogeneity in causal questions which can be particularly informative for the efficient allocation of public resources if sub-populations react differently to a specific policy measure. These advances show that besides the power of statistical learning in prediction, it has also led to important advances in causal inference. This provides yet a further argument for economists to integrate these methods into their toolkit.

2. At several stages of this thesis, it has been highlighted that an important role of economic research is to assist the policy decision-making process. Just in a similar vein as industry turns data into profitable business models, policy can also benefit from effectively exploiting information from increasingly growing sources of digital resources. This study has presented two prototypes of such policy tools: (1) a real-time decision support framework in times of economic shocks in Chapter 2.2, (2) a technology mapping framework for steering technological change more effectively in Chapter 3.1. Developing such prototypes into operational systems may help in reducing red tape and profligacy when spending public resources. A good example provides the second paper of this

[b] Arguably, the estimation of sector-size-specific insolvency gaps in the first paper in Chapter 2 could be interpreted as a causal consequence of the COVID-19 induced policy intervention. However, there are two reasons why I refrain from this causal interpretation. First, the estimation of the counterfactual outcome does rely on a control group in the conventional sense. Due to the indiscriminate granting of aid measures, it was not possible to construct *contemporaneous* control groups. Instead, we needed to match control units based on credit rating changes observed prior to the pandemic. Second, COVID-19 induced a wide range of distinct policy responses whose effects on corporate survival cannot be isolated in our research design. So it can only be argued that the *overall* policy response has favored a backlog of insolvency filings among small firms.

thesis where timely dissemination of information to policymakers may allow them to offer liquidity support more selectively. This could help avoiding subsectors or individual companies to absorb support measures as windfall gains. Clearly, the development of such information support systems needs to be further explored to realize their full potential for policymakers.

Besides the opportunities that such policy instruments offer, they are not without risks. In fact, data-driven decision support can lead to systematic discrimination. By using sensitive variables that affect protected groups (e.g. gender, religion, income), there is the danger that such decision support systems (un)intentionally produce biased recommendations at the expense of minority groups. Even if the sensitive variables themselves are not included in the model development, there is still a risk of discrimination, as biases may arise from other covariates that are highly correlated with one of the sensitive variables (Barredo Arrieta et al., 2020). So with artificial intelligence becoming increasingly more widespread in political decision-making processes, major challenges arise for research. From a methodological perspective, approaches are needed that make the predictions of computer-aided decision-making systems explainable and robust against systematic discrimination. Explainable Artificial Intelligence (XAI) is a young stream of research that aims at developing such systems. But also from a philosophical point of view, the question increasingly comes up where to draw the line in liberal democracies in the transfer of political decisions from the representatives of the people to data-driven decision-making systems. Although answering these questions is certainly beyond the scope of this thesis, it is important to consider them in developing prototypes for data-driven policy advisory tools, especially in a world where digitization and big data becomes ever more omnipresent.

Appendices

Appendix A

Small Firms and the COVID-19 Insolvency Gap

Small Bus Econ (2022) 58:887–917
https://doi.org/10.1007/s11187-021-00514-4

Small firms and the COVID-19 insolvency gap

**Julian Oliver Dörr · Georg Licht ·
Simona Murmann**

Accepted: 12 May 2021 / Published online: 6 July 2021
© The Author(s) 2021

Abstract COVID-19 placed a special role on fiscal policy in rescuing companies short of liquidity from insolvency. In the first months of the crisis, SMEs as the backbone of Germany's economy benefited from large and mainly indiscriminate aid measures. Avoiding business failures in a whatever-it-takes fashion contrasts, however, with the cleansing mechanism of economic crises: a mechanism which forces unviable firms out of the market, thereby reallocating resources efficiently. By focusing on firms' pre-crisis financial standing, we estimate the extent to which the policy response induced an insolvency gap and analyze whether the gap is characterized by firms which were already struggling before the pandemic. With the policy measures being focused on smaller firms, we also examine whether this insolvency gap differs with respect to firm size. Our results show that the COVID-19 policy response in Germany has triggered a backlog of insolvencies that is particularly pronounced among financially weak, small firms, having potential long-term implications on entrepreneurship and economic recovery.

Plain English Summary This study analyzes the extent to which the strong policy support to companies in the early phase of the COVID-19 crisis has prevented a large wave of corporate insolvencies. Using data of about 1.5 million German companies, it is shown that it was mainly smaller firms that experienced strong financial distress and would have gone bankrupt without policy assistance. In times of crises, insolvencies usually allow for a reallocation of employees and capital to more efficient firms. However, the analysis reveals that this 'cleansing effect' is hampered in the current crisis as the largely indiscriminate granting of liquidity subsidies and the temporary suspension of the duty to file for insolvency have caused an insolvency gap that is driven by firms which were already in a weak financial position before the crisis. Overall, the insolvency gap is estimated to affect around 25,000 companies, a substantial number compared to the around 16,300 actual insolvencies in 2020. In the ongoing crisis, policy makers should prefer instruments favoring entrepreneurs who respond innovatively to the pandemic instead of prolonging the survival of near-insolvent firms.

Keywords COVID-19 · Liquidity support · Insolvency law · Corporate insolvency · Cleansing effect · SMEs · Entrepreneurship

JEL Classifications G33 · H12 · H81 · L26

J. O. Dörr (✉) · G. Licht · S. Murmann
Department of Economics of Innovation and Industrial Dynamics, ZEW Leibniz Centre for European Economic Research, L7 1, 68161 Mannheim, Germany
e-mail: julian.doerr@zew.de

J. O. Dörr
Department of Econometrics and Statistics,
Justus Liebig University Giessen, Licher Straße 64,
35394 Gießen, Germany

1 Introduction

COVID-19 and its unprecedented economic impacts have ground economies worldwide to a halt. As a result of the early lockdown measures to contain the spread of the virus, many companies faced reduced business activity and declining sales, which had an immediate impact on their liquidity positions. Indeed, both the negative demand shock paired with a negative supply shock in most industries have put numerous companies under severe pressure to keep their operations afloat. Previous crises have taught that small entrepreneurial firms are particularly prone to considerable liquidity constraints in deep recessions. For example, literature on the financial crisis of 2007–2009 shows that especially small and entrepreneurial enterprises were exposed to a severe liquidity crunch due to the collapse of the interbank market and its negative impact on corporate lending (e.g. Cowling et al. 2012; Iyer et al. 2014; Lee et al. 2015; McGuinness & Hogan 2016). In the COVID-19 crisis, the early effects of the combined negative supply and demand shock are also characterized by a deep liquidity shock in the real economy. The decline of trading activities and lack of business revenues made many firms dependent on their cash reserves in order to meet their unchanged fixed cost obligations. As smaller and entrepreneurial companies are characterized by strong dependence on internally generated funds to capitalize their business and provide the liquidity needed to finance day-to-day operations, both their cash reserves and collaterals for external financing are generally limited (Cowling et al., 2020). In times of financial distress as in the current COVID-19 crisis, this makes small ventures particularly vulnerable to financial insolvency (Fairlie, 2020; Bartik et al., 2020). Recent research suggests that severely affected small entrepreneurial ventures even seek for alternative financing methods such as bootstrap financing to keep their businesses alive (Block et al., 2021). Trapped in a situation of thin capital reserves and lack of collaterals for drawing new credit lines, small firms face therefore a particularly high risk of business failure without the relief through policy intervention.

Aware of the far-reaching consequences of a wave of corporate insolvencies, governments in almost all countries have initiated a series of emergency measures to strengthen liquidity positions of their national companies, some of which exclusively focusing on easing the burden of Small and Medium-sized Enterprises (SMEs) (OECD, 2020a). In the European Union (EU), for instance, member states' liquidity support in form of public loan guarantees and tax deferrals for distressed sectors has increased by an estimated 6 percentage points (pp) of EU GDP compared to pre-crisis levels (Council of the European Union, 2020). In most countries, policy measures have gone beyond deferrals and loan guarantees, including instruments such as wage subsidies and adjustments in bankruptcy regimes. While there is no doubt that a strong policy response was necessary to keep the struggling economy afloat, the need to respond quickly and the sheer volume of firms seeking assistance left little time for policymakers to assess the viability of firms that received early government support. Thus, many of the early policy measures were not only unprecedented in scale but also largely granted indiscriminately with the primary focus to avoid corporate bankruptcies.[1] Even though some programs' eligibility criteria are formally linked to pandemic-induced financial distress only, information asymmetries make drawing a line often difficult in reality. We argue that these circumstances have favored a substantial backlog of corporate insolvencies as policy measures have also kept otherwise insolvent firms in the market. This phenomenon is referred to as insolvency gap in the remaining of the paper.

The central purpose of this study is to analyze whether the early policy response has indeed induced such an insolvency gap and, if so, by which firms the gap is mainly driven. We do so by incorporating the Schumpeterian cleansing effect usually observed in economic crises into our analysis. In Schumpeterian economics, crises are typically seen as cleansing mechanism forcing unviable firms out of the market thereby efficiently reallocating resources to more productive companies. Our hypothesis is that this cleansing mechanism is strongly compromised by the undifferentiated policy response which favors the survival of otherwise unviable firms. Since in times of crises

[1] In Germany, for instance, liquidity grants' 'application and payment process is to be swift and free from red tape' according to the Ministry of Finance (Federal Ministry of Finance 2020b, para. 2). Moreover, in context of public loan programs it is stated that 'the credit approval process does not involve additional credit risk assessment by the bank' and that 'there are no requirements for collateral security' (Federal Ministry for Economic Affairs and Energy 2020, para. 5).

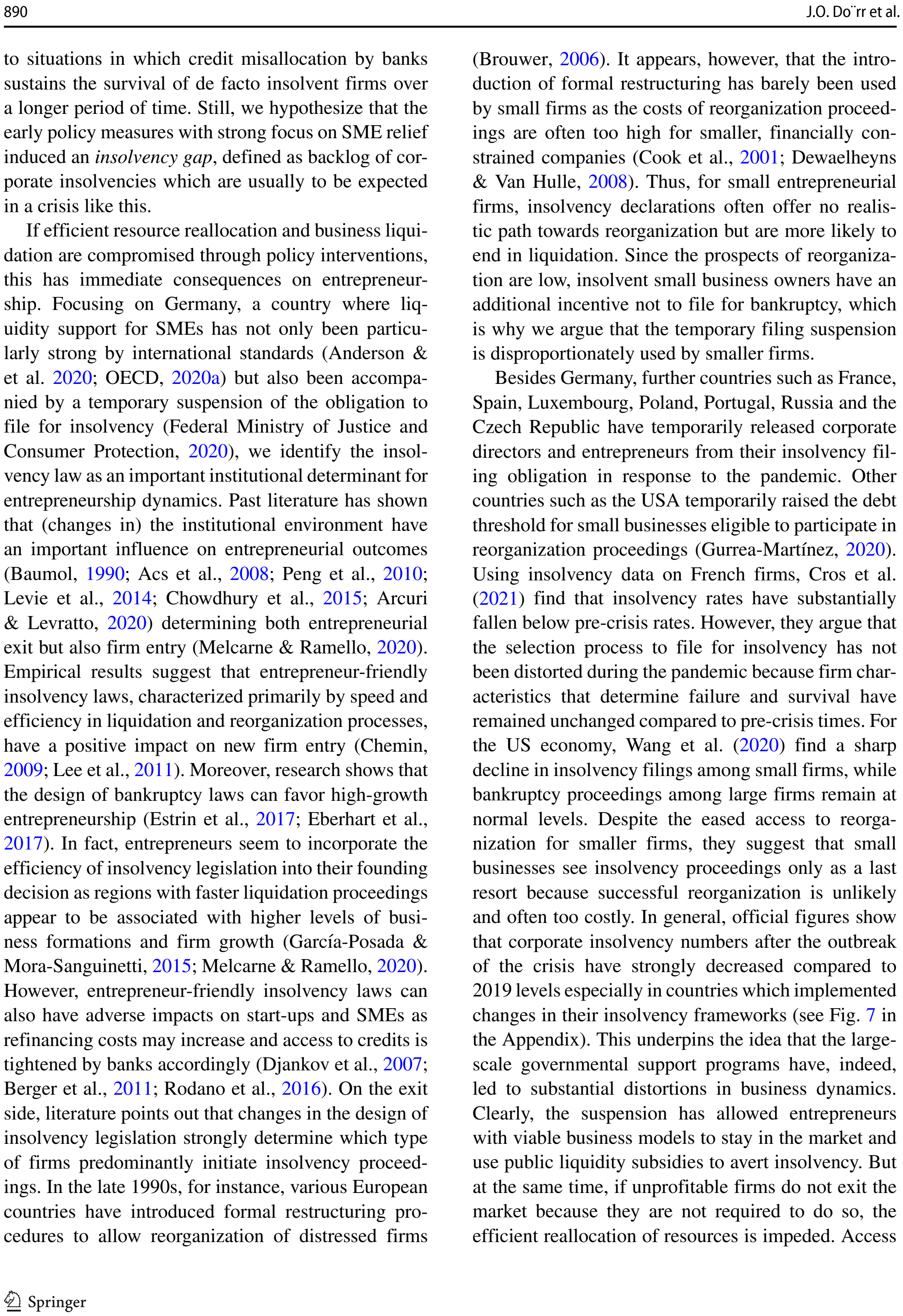
small firms tend to be particularly prone to liquidity shortages, we believe that the risk of unviable 'survivors' is especially high among smaller enterprises. The strong policy focus on SMEs in many countries (OECD, 2020a) reinforces this hypothesis. Finally, it is likely that the prolonged expansion prior to COVID-19 along with the low interest rate environment have already accumulated a substantial number of financially weak companies before the pandemic (Barrero et al., 2020). Normally, the COVID-19 crisis would have been a 'natural' mechanism to force such ailing firms out of the market. Given the interplay between prolonged expansion and sudden economic decline paired with a strong policy response, our hypothesis is that the insolvency gap is strongly driven by small firms with weak financial conditions prior to the crisis.

Our contribution to the fast growing literature on the economic effects of the COVID-19 crisis is manifold. First, we examine the heterogeneity with respect to firm size in policy makers' response to the risk of large-scale business failures. Second, we translate Schumpeter's theory of the cleansing effect in economic crises into an empirical assessment by estimating the size of a policy-induced insolvency gap using firm-specific credit rating data combined with information on insolvency filings. Controlling for updates in a firm's credit rating, we estimate the insolvency gap induced by the COVID-19-related policy measures using a potential outcome setting. Based on pre-crisis observations of no policy intervention comparable firms with closely matching changes in their credit rating are used as control group for the estimation of counterfactual insolvency rates. Finally, we discuss the consequences for entrepreneurship if efficient resource reallocation and business liquidation are compromised.

The remainder of the paper proceeds as follows. Section 2 gives an overview of the relevant literature. In Section 3, we discuss the different COVID-19 support instruments for firms in Germany and emphasize their different orientations depending on firm size. Section 4 introduces the data sources and variables used to estimate the insolvency gap. Moreover, the framework for the matching of counterfactual survival states is introduced. Section 5 empirically examines the adverse impacts of the pandemic and its heterogeneity across firms of different size and sector affiliation. Moreover, it presents the empirical results of the insolvency gap estimation. Ultimately, Section 6 discusses the implications of our results and concludes.

2 Related literature

The fast growing literature on business failures in response to the adverse economic impacts of COVID-19 stresses that the early assistance packages may bare high economic costs if they keep unviable firms alive (Kalemli-Ozcan et al., 2020; Barrero et al., 2020; Cowling et al., 2020; Juergensen et al., 2020; OECD, 2020b; Didier et al., 2021). Kalemli-Ozcan et al. (2020), for example, find for a number of European countries that without appropriate targeting of policy instruments, the fiscal costs of intervention and the number of 'ghost' firms kept alive are substantially higher compared to a scenario in which policies target only 'viable' firms. Besides the direct fiscal costs associated with indiscriminate policy interventions, there is yet another source of economic costs associated with keeping unviable firms alive. In Schumpeterian economics, this may also impede the cleansing effect of creative destruction (see, for example, Legrand 2017 and in the COVID-19 context Barrero et al. 2020; Guerini et al. 2020). This effect describes a process in which resources are reallocated from less efficient and less creative firms to more efficient ones enhancing overall economic productivity and innovation (Schumpeter, 1942). Typically, this process of efficient resource reallocation is particularly strong in times of economic crisis, allowing viable and innovative firms to gain market share as unprofitable firms exit the market (Caballero and Hammour, 1994; Archibugi et al., 2013; Carreira & Teixeira, 2016). As such, without the intervention of fiscal policy, business failures of unviable firms are expected to be substantial in economic recessions and, given the strong vulnerability of small and entrepreneurial firms, the effect is expected to be particularly pronounced among smaller businesses. In the current crisis, however, there is growing public concern that this process of creative destruction and 'cleanse out' of unviable firms is seriously hampered by an increasing policy-induced 'zombification' of the economy (see, e.g., The Economist 2020a; The Washington Post 2020). Analyzing only the short-term effects of policy aid on firm survival, we do not want to go as far as speaking of a zombification which typically refers



to situations in which credit misallocation by banks sustains the survival of de facto insolvent firms over a longer period of time. Still, we hypothesize that the early policy measures with strong focus on SME relief induced an *insolvency gap*, defined as backlog of corporate insolvencies which are usually to be expected in a crisis like this.

If efficient resource reallocation and business liquidation are compromised through policy interventions, this has immediate consequences on entrepreneurship. Focusing on Germany, a country where liquidity support for SMEs has not only been particularly strong by international standards (Anderson & et al. 2020; OECD, 2020a) but also been accompanied by a temporary suspension of the obligation to file for insolvency (Federal Ministry of Justice and Consumer Protection, 2020), we identify the insolvency law as an important institutional determinant for entrepreneurship dynamics. Past literature has shown that (changes in) the institutional environment have an important influence on entrepreneurial outcomes (Baumol, 1990; Acs et al., 2008; Peng et al., 2010; Levie et al., 2014; Chowdhury et al., 2015; Arcuri & Levratto, 2020) determining both entrepreneurial exit but also firm entry (Melcarne & Ramello, 2020). Empirical results suggest that entrepreneur-friendly insolvency laws, characterized primarily by speed and efficiency in liquidation and reorganization processes, have a positive impact on new firm entry (Chemin, 2009; Lee et al., 2011). Moreover, research shows that the design of bankruptcy laws can favor high-growth entrepreneurship (Estrin et al., 2017; Eberhart et al., 2017). In fact, entrepreneurs seem to incorporate the efficiency of insolvency legislation into their founding decision as regions with faster liquidation proceedings appear to be associated with higher levels of business formations and firm growth (García-Posada & Mora-Sanguinetti, 2015; Melcarne & Ramello, 2020). However, entrepreneur-friendly insolvency laws can also have adverse impacts on start-ups and SMEs as refinancing costs may increase and access to credits is tightened by banks accordingly (Djankov et al., 2007; Berger et al., 2011; Rodano et al., 2016). On the exit side, literature points out that changes in the design of insolvency legislation strongly determine which type of firms predominantly initiate insolvency proceedings. In the late 1990s, for instance, various European countries have introduced formal restructuring procedures to allow reorganization of distressed firms (Brouwer, 2006). It appears, however, that the introduction of formal restructuring has barely been used by small firms as the costs of reorganization proceedings are often too high for smaller, financially constrained companies (Cook et al., 2001; Dewaelheyns & Van Hulle, 2008). Thus, for small entrepreneurial firms, insolvency declarations often offer no realistic path towards reorganization but are more likely to end in liquidation. Since the prospects of reorganization are low, insolvent small business owners have an additional incentive not to file for bankruptcy, which is why we argue that the temporary filing suspension is disproportionately used by smaller firms.

Besides Germany, further countries such as France, Spain, Luxembourg, Poland, Portugal, Russia and the Czech Republic have temporarily released corporate directors and entrepreneurs from their insolvency filing obligation in response to the pandemic. Other countries such as the USA temporarily raised the debt threshold for small businesses eligible to participate in reorganization proceedings (Gurrea-Martínez, 2020). Using insolvency data on French firms, Cros et al. (2021) find that insolvency rates have substantially fallen below pre-crisis rates. However, they argue that the selection process to file for insolvency has not been distorted during the pandemic because firm characteristics that determine failure and survival have remained unchanged compared to pre-crisis times. For the US economy, Wang et al. (2020) find a sharp decline in insolvency filings among small firms, while bankruptcy proceedings among large firms remain at normal levels. Despite the eased access to reorganization for smaller firms, they suggest that small businesses see insolvency proceedings only as a last resort because successful reorganization is unlikely and often too costly. In general, official figures show that corporate insolvency numbers after the outbreak of the crisis have strongly decreased compared to 2019 levels especially in countries which implemented changes in their insolvency frameworks (see Fig. 7 in the Appendix). This underpins the idea that the large-scale governmental support programs have, indeed, led to substantial distortions in business dynamics. Clearly, the suspension has allowed entrepreneurs with viable business models to stay in the market and use public liquidity subsidies to avert insolvency. But at the same time, if unprofitable firms do not exit the market because they are not required to do so, the efficient reallocation of resources is impeded. Access to skilled human capital, physical resources such as office space, and bank loans is limited for newly entering firms when unviable firms congest the market. This may prevent future entrepreneurs from starting up, but can also discourage existing entrepreneurs from initiating new, more promising ventures. Hence, we assume that an insolvency gap along with a further prolongation of aid measures in the ongoing crises is likely to result in a decrease in entrepreneurial activity in the longer term.

3 Policy response in Germany

Official figures show that in Germany, the fiscal policy response to prevent corporate insolvencies due to crisis-related liquidity bottlenecks is particularly pronounced by international comparison. According to a comparative study of the economic think tank Bruegel, nearly 40% of Germany's 2019 GDP was spent on COVID-19 measures to strengthen companies' liquidity positions (Anderson and et al. 2020). Compared with a number of selected OECD countries, this is the second strongest response in terms of a country's overall economic performance (see Fig. 6 in the Appendix). In fact, the German Federal Government itself describes the response as the 'largest assistance package in the history of the Federal Republic of Germany' (Federal Ministry of Finance 2020d, 3).

From a small business economics view, it is interesting to see that a number of intervention measures adopted by the German Federal Government have been specifically designed to target SMEs (OECD, 2020a). In the following, we describe the policy instruments to counter the economic impacts of the COVID-19 crisis in more detail, focusing on how the instruments differ with respect to firm size (for a quick overview of the measures and possible effects on corporate insolvencies the reader is referred to Table 10 in the Appendix).

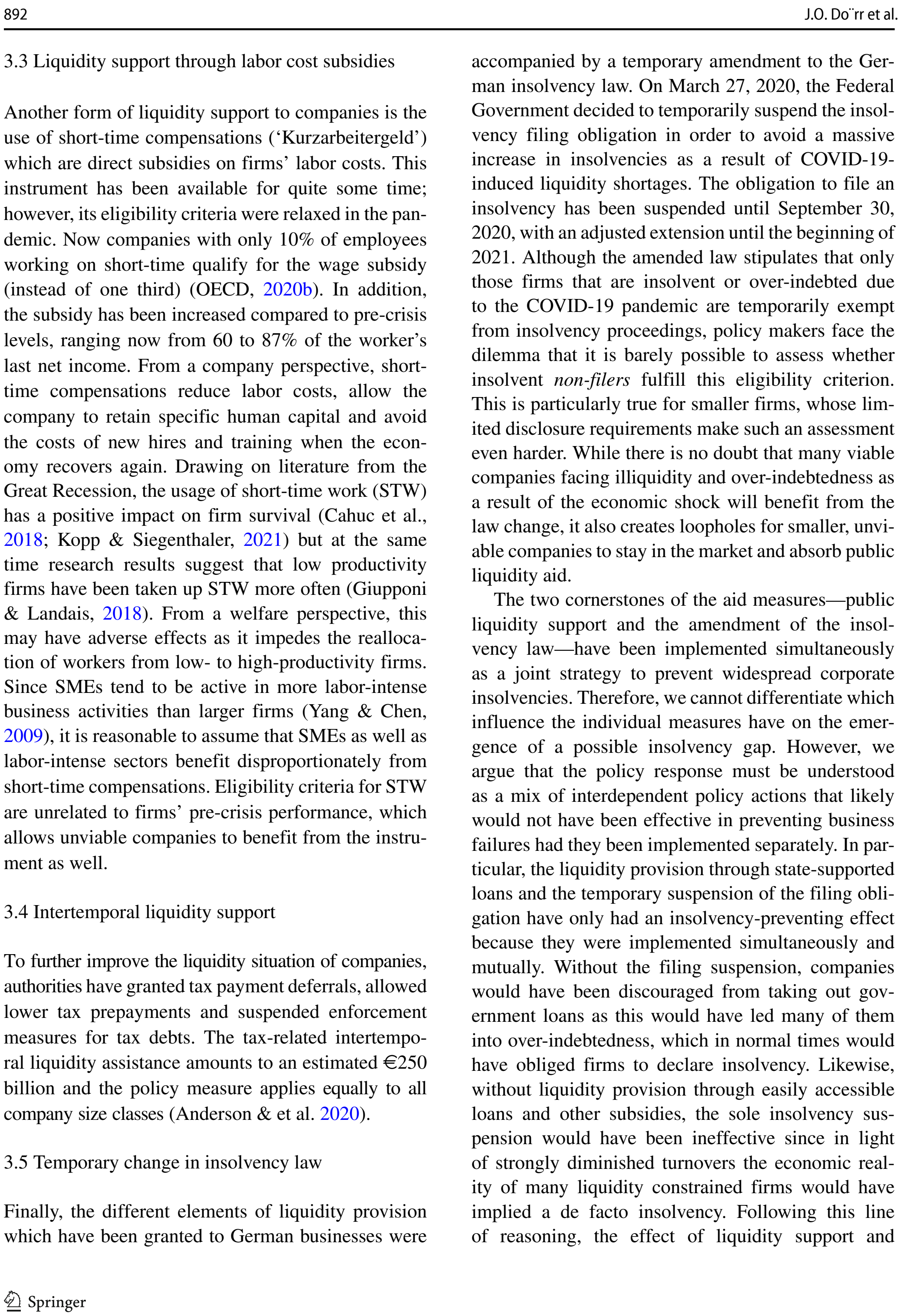
3.1 Direct liquidity subsidies

As an immediate response to the first lockdown, the Federal Government granted liquidity subsidies through direct cash transfers ('Sofort-' and 'Überbrückungshilfen'). The extent of liquidity support is primarily determined by company size, measured by the number of employees or previous revenues. In case of the 'Soforthilfen', for instance, only micro-firms with up to 10 employees were eligible to receive injections between €9,000 and €15,000 for three months to cover their operational costs (Federal Ministry of Finance, 2020d). These immediate subsidies have been accompanied by a large-scale stimulus package worth €25 billion covering a substantial part of SMEs' fixed operating costs (Federal Ministry of Finance, 2020c). Generally, the subsidies were granted in a non-bureaucratic fashion easily accessible to all micro-businesses and SMEs which assured that they were suffering financial distress because of the COVID-19 pandemic (Federal Government of Germany, 2020).

3.2 Liquidity loans under public guarantee schemes

For SMEs with more than 10 employees the KfW Instant Loan Program has been launched. The program offers SMEs loans that are fully collateralized by the state. These loans amount up to 25% of a firm's 2019 revenues with a cap of €500k for small companies and €800k for medium-sized companies, respectively. No credit risk assessments are taking place and no collaterals are required. The only eligibility criterion is that the company was profitable in 2019 or at least on average profitable between 2017 and 2019 (Federal Ministry for Economic Affairs and Energy, 2020). This fairly broad criterion shows that the process is focused on speed and ease applied 'without red tape' (Federal Ministry of Finance 2020b, 1) and not on elaborate screening mechanisms that could prevent providing liquidity to unviable firms.

Furthermore, the COVID-19 support package includes additional government guarantees on loans for both small and larger businesses, including lower interest rates for small firms compared to large firms. (Federal Ministry for Economic Affairs and Energy, 2020). Similar to the Instant Loan Program, the loans are channeled through commercial banks and the state-owned bank KfW assumes risk coverage of 80% for large enterprises and 90% for SMEs with a simplified risk assessment (Federal Ministry of Finance, 2020a). For commercial banks, this makes lending to SMEs particularly attractive and, given that they only bear 10% of the default risk, further disincentivizes comprehensive risk assessments by the issuing bank.



3.3 Liquidity support through labor cost subsidies

Another form of liquidity support to companies is the use of short-time compensations ('Kurzarbeitergeld') which are direct subsidies on firms' labor costs. This instrument has been available for quite some time; however, its eligibility criteria were relaxed in the pandemic. Now companies with only 10% of employees working on short-time qualify for the wage subsidy (instead of one third) (OECD, 2020b). In addition, the subsidy has been increased compared to pre-crisis levels, ranging now from 60 to 87% of the worker's last net income. From a company perspective, short-time compensations reduce labor costs, allow the company to retain specific human capital and avoid the costs of new hires and training when the economy recovers again. Drawing on literature from the Great Recession, the usage of short-time work (STW) has a positive impact on firm survival (Cahuc et al., 2018; Kopp & Siegenthaler, 2021) but at the same time research results suggest that low productivity firms have been taken up STW more often (Giupponi & Landais, 2018). From a welfare perspective, this may have adverse effects as it impedes the reallocation of workers from low- to high-productivity firms. Since SMEs tend to be active in more labor-intense business activities than larger firms (Yang & Chen, 2009), it is reasonable to assume that SMEs as well as labor-intense sectors benefit disproportionately from short-time compensations. Eligibility criteria for STW are unrelated to firms' pre-crisis performance, which allows unviable companies to benefit from the instrument as well.

3.4 Intertemporal liquidity support

To further improve the liquidity situation of companies, authorities have granted tax payment deferrals, allowed lower tax prepayments and suspended enforcement measures for tax debts. The tax-related intertemporal liquidity assistance amounts to an estimated €250 billion and the policy measure applies equally to all company size classes (Anderson & et al. 2020).

3.5 Temporary change in insolvency law

Finally, the different elements of liquidity provision which have been granted to German businesses were accompanied by a temporary amendment to the German insolvency law. On March 27, 2020, the Federal Government decided to temporarily suspend the insolvency filing obligation in order to avoid a massive increase in insolvencies as a result of COVID-19-induced liquidity shortages. The obligation to file an insolvency has been suspended until September 30, 2020, with an adjusted extension until the beginning of 2021. Although the amended law stipulates that only those firms that are insolvent or over-indebted due to the COVID-19 pandemic are temporarily exempt from insolvency proceedings, policy makers face the dilemma that it is barely possible to assess whether insolvent *non-filers* fulfill this eligibility criterion. This is particularly true for smaller firms, whose limited disclosure requirements make such an assessment even harder. While there is no doubt that many viable companies facing illiquidity and over-indebtedness as a result of the economic shock will benefit from the law change, it also creates loopholes for smaller, unviable companies to stay in the market and absorb public liquidity aid.

The two cornerstones of the aid measures—public liquidity support and the amendment of the insolvency law—have been implemented simultaneously as a joint strategy to prevent widespread corporate insolvencies. Therefore, we cannot differentiate which influence the individual measures have on the emergence of a possible insolvency gap. However, we argue that the policy response must be understood as a mix of interdependent policy actions that likely would not have been effective in preventing business failures had they been implemented separately. In particular, the liquidity provision through state-supported loans and the temporary suspension of the filing obligation have only had an insolvency-preventing effect because they were implemented simultaneously and mutually. Without the filing suspension, companies would have been discouraged from taking out government loans as this would have led many of them into over-indebtedness, which in normal times would have obliged firms to declare insolvency. Likewise, without liquidity provision through easily accessible loans and other subsidies, the sole insolvency suspension would have been ineffective since in light of strongly diminished turnovers the economic reality of many liquidity constrained firms would have implied a de facto insolvency. Following this line of reasoning, the effect of liquidity support and temporary change in insolvency law on (non-) selection into insolvency is best analyzed as a policy mix used to combat the threat of mass insolvencies. While the insolvency filing suspension allowed both small and large companies to avert insolvency and possibly survive the crisis by taking advantage of liquidity injections by the state (Federal Ministry of Justice and Consumer Protection, 2020), it has been shown that many of the liquidity support measures directly target SMEs or provide indirect channels for especially smaller (and often entrepreneurial) businesses to benefit disproportionately. With few screening mechanisms in place, there is the risk that unviable firms will be kept alive, freezing up resources that could be used more productively elsewhere and possibly hampering entrepreneurial activity.

This section has highlighted the role of policy support to counter the economic consequences of the pandemic in Germany—a country that has provided substantial assistance to businesses to avoid a wave of corporate bankruptcies. It has suggested that the joint implementation of widespread but undifferentiated liquidity support strongly focusing on SMEs together with the temporary amendment of the insolvency law, is likely to have favored a backlog of corporate insolvencies particularly pronounced among small and possibly financially weak companies. In the next section, we introduce the data and methodology we use to estimate the existence and extent of such an insolvency gap.

4 Data, variables and methodology

4.1 Data and variables

The study uses two data sources which both originate from the Mannheim Enterprise Panel (MUP) covering the near universe of economically active firms in Germany (Bersch et al., 2014). The first data source is a survey where the questioned companies have been sampled from the MUP. The survey is used to examine how companies of different size and in different sectors are affected by the adverse impacts of the COVID-19 pandemic and motivates why we estimate the insolvency gap distinguishing between sector affiliation and company size. For the estimation of the insolvency gap we use a second data source: a large sample of firm-specific credit rating information along with information concerning the firms' insolvency status. In the following, we will introduce both data sources and the variables used in this study in more detail.

4.1.1 Survey data

We employ the survey to primarily assess which industries and company sizes are affected most by the crisis. Based on a representative random sample of German companies, drawn from the MUP and stratified by firm size and industry affiliation the survey was conducted three times spanning the period in which the German insolvency regime was fully suspended.[2] The survey includes questions on COVID-19-related economic effects on various business dimensions. The collected data has then been supplemented with credit rating scores from the MUP, which allows to control for the financial situation of the companies prior to the crisis. As shown in Fig. 1, we use the survey data to investigate whether the adverse economic impacts of COVID-19 differ across sectors and firm size classes. These results together with the heterogeneity in public aid programs with respect to firm size as outlined in Section 3 motivates us to conduct our main empirical analysis at the sector-size level.

Table 1 shows summary statistics of the relevant variables used to construct a COVID-19 Exposure Index, CEI, reflecting the extent to which firms experienced negative impacts in relation to the pandemic. Firms were asked on a Lickert scale of 0 to 4 in which of the following areas they experienced negative impacts as a result of the COVID-19 crisis: (1) decrease in demand, (2) shutdown of production, (3) supply chain interruption, (4) staffing shortage, (5) logistical difficulties, (6) liquidity shortfalls.[3] From these six questions we construct CEI as simple sum of the response values. The average index is 6.31 out of a maximum possible value of 24. The most common and most severe impact relates to the decline in demand, where respondents reported an average negative impact of 1.85. Shutdown of production facilities

[2] The surveys have been conducted in April 2020, in June 2020 and in September 2020 spanning the period of the full suspension of the obligation to file for insolvency and is therefore particularly suitable for capturing the early policy-induced effects of the crisis.

[3] 0 indicates no negative effects, 4 signals strong negative effects.

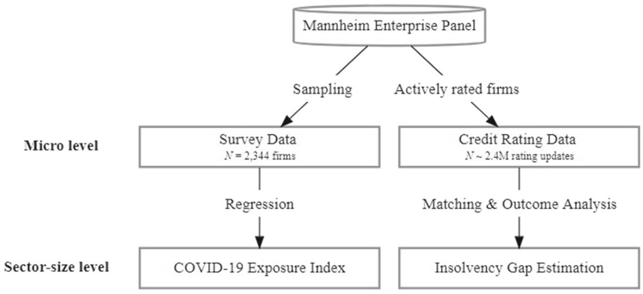


Fig. 1 Data sources used in this study. Note: Observations of the survey data (companies) and credit rating data (firm-specific rating revisions) originate from the Mannheim Enterprise Panel (MUP) data base. Survey data allows to estimate exposure to adverse effects of the pandemic on the sector-size level. Credit rating data is used to estimate the existence of an insolvency gap on the sector-size level

and liquidity bottlenecks are also frequently mentioned consequences.

4.1.2 Credit rating data

For the purpose of estimating whether the bankruptcy filing behavior has changed significantly as a result of the crisis-related aid measures and possibly created a backlog of insolvencies, we examine credit rating updates of close to all economically active firms listed in the MUP.[4] The Mannheim Enterprise Panel is particularly suited for an analysis of the insolvency-related cleansing effect as it is constructed by processing and structuring data collected by Creditreform, the leading credit agency in Germany. Creditreform regularly measures and updates the creditworthiness of German companies. Overall our sample comprises 2,373,782 credit rating updates of 1,500,764 distinct German businesses whose ratings were updated at least once during the last three years.[5] Table 2 shows that the sample of about 1.5 million companies is very diverse in its industry and size composition. Most important in the context of this study is the coverage of SMEs, which not only is representative for the German economy (see Table 2), but also allows for a nuanced differentiation between medium-sized, small and micro-enterprises. Therefore, it suits well to examine the policy-induced heterogeneity of the COVID-19 related effects on business failures with a special focus on possible size differences not only among SMEs and large enterprises but also within the group of SMEs. The latter estimation of the insolvency gap will be conducted on the sector-size level as displayed in Table 2.

Assuming that the COVID-19 shock and its economic consequences on liquidity and insolvency distress of German businesses began by the end of March 2020, we split our sample into a 'pre-crisis' period and a 'crisis' period. This cutoff point also captures COVID-19 policy dynamics as the German government imposed the first countrywide lockdown that includes a shutdown of most customer service-related businesses on March 22 and suspended the obligation to file for bankruptcy on March 27 (Federal Ministry of Justice and Consumer Protection, 2020). Consequently, the pre-crisis period comprises all credit rating updates which took place between July 2017 and December 2019. The crisis period includes all observations between April 2020 and end of July 2020.[6] In the later estimation of the insolvency gap, rating updates from the pre-crisis period serve as pool of control observations. Closely matching credit rating updates from this pool are used to estimate counterfactual insolvency rates which will be compared against

[4] In our analysis a company is defined as economically active if it has received a credit rating update at least once over the last three years spanning the period from July 2017 to July 2020.

[5] We observe one and the same company at most three times in our sample. Thus, credit rating updates normally do not take place more often than once per year but may be conducted in a less regular cycle.

[6] Note that we exclude observations between January 2020 and March 2020 which we see as transitional phase in which assignment to either of the two periods is not straightforward. Also note that July 2020 is the latest month for which we observe credit rating information.

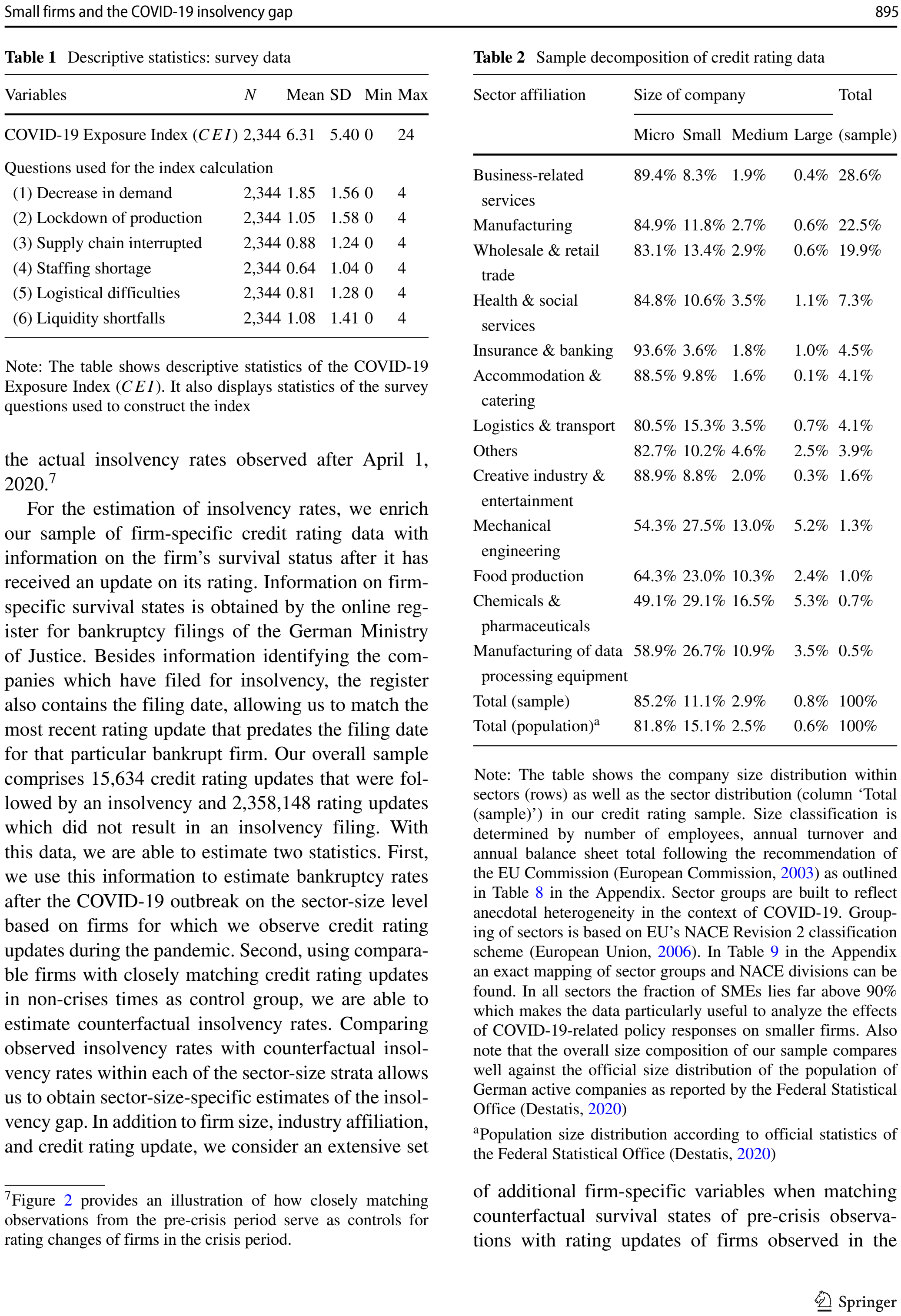


Table 1 Descriptive statistics: survey data

Variables	*N*	Mean	SD	Min	Max
COVID-19 Exposure Index (CEI)	2,344	6.31	5.40	0	24
Questions used for the index calculation					
(1) Decrease in demand	2,344	1.85	1.56	0	4
(2) Lockdown of production	2,344	1.05	1.58	0	4
(3) Supply chain interrupted	2,344	0.88	1.24	0	4
(4) Staffing shortage	2,344	0.64	1.04	0	4
(5) Logistical difficulties	2,344	0.81	1.28	0	4
(6) Liquidity shortfalls	2,344	1.08	1.41	0	4

Note: The table shows descriptive statistics of the COVID-19 Exposure Index (CEI). It also displays statistics of the survey questions used to construct the index

Table 2 Sample decomposition of credit rating data

Sector affiliation	Size of company				Total
	Micro	Small	Medium	Large	(sample)
Business-related services	89.4%	8.3%	1.9%	0.4%	28.6%
Manufacturing	84.9%	11.8%	2.7%	0.6%	22.5%
Wholesale & retail trade	83.1%	13.4%	2.9%	0.6%	19.9%
Health & social services	84.8%	10.6%	3.5%	1.1%	7.3%
Insurance & banking	93.6%	3.6%	1.8%	1.0%	4.5%
Accommodation & catering	88.5%	9.8%	1.6%	0.1%	4.1%
Logistics & transport	80.5%	15.3%	3.5%	0.7%	4.1%
Others	82.7%	10.2%	4.6%	2.5%	3.9%
Creative industry & entertainment	88.9%	8.8%	2.0%	0.3%	1.6%
Mechanical engineering	54.3%	27.5%	13.0%	5.2%	1.3%
Food production	64.3%	23.0%	10.3%	2.4%	1.0%
Chemicals & pharmaceuticals	49.1%	29.1%	16.5%	5.3%	0.7%
Manufacturing of data processing equipment	58.9%	26.7%	10.9%	3.5%	0.5%
Total (sample)	85.2%	11.1%	2.9%	0.8%	100%
Total (population)[a]	81.8%	15.1%	2.5%	0.6%	100%

Note: The table shows the company size distribution within sectors (rows) as well as the sector distribution (column ‘Total (sample)’) in our credit rating sample. Size classification is determined by number of employees, annual turnover and annual balance sheet total following the recommendation of the EU Commission (European Commission, 2003) as outlined in Table 8 in the Appendix. Sector groups are built to reflect anecdotal heterogeneity in the context of COVID-19. Grouping of sectors is based on EU’s NACE Revision 2 classification scheme (European Union, 2006). In Table 9 in the Appendix an exact mapping of sector groups and NACE divisions can be found. In all sectors the fraction of SMEs lies far above 90% which makes the data particularly useful to analyze the effects of COVID-19-related policy responses on smaller firms. Also note that the overall size composition of our sample compares well against the official size distribution of the population of German active companies as reported by the Federal Statistical Office (Destatis, 2020)

[a]Population size distribution according to official statistics of the Federal Statistical Office (Destatis, 2020)

the actual insolvency rates observed after April 1, 2020.[7]

For the estimation of insolvency rates, we enrich our sample of firm-specific credit rating data with information on the firm’s survival status after it has received an update on its rating. Information on firm-specific survival states is obtained by the online register for bankruptcy filings of the German Ministry of Justice. Besides information identifying the companies which have filed for insolvency, the register also contains the filing date, allowing us to match the most recent rating update that predates the filing date for that particular bankrupt firm. Our overall sample comprises 15,634 credit rating updates that were followed by an insolvency and 2,358,148 rating updates which did not result in an insolvency filing. With this data, we are able to estimate two statistics. First, we use this information to estimate bankruptcy rates after the COVID-19 outbreak on the sector-size level based on firms for which we observe credit rating updates during the pandemic. Second, using comparable firms with closely matching credit rating updates in non-crises times as control group, we are able to estimate counterfactual insolvency rates. Comparing observed insolvency rates with counterfactual insolvency rates within each of the sector-size strata allows us to obtain sector-size-specific estimates of the insolvency gap. In addition to firm size, industry affiliation, and credit rating update, we consider an extensive set of additional firm-specific variables when matching counterfactual survival states of pre-crisis observations with rating updates of firms observed in the

[7]Figure 2 provides an illustration of how closely matching observations from the pre-crisis period serve as controls for rating changes of firms in the crisis period.

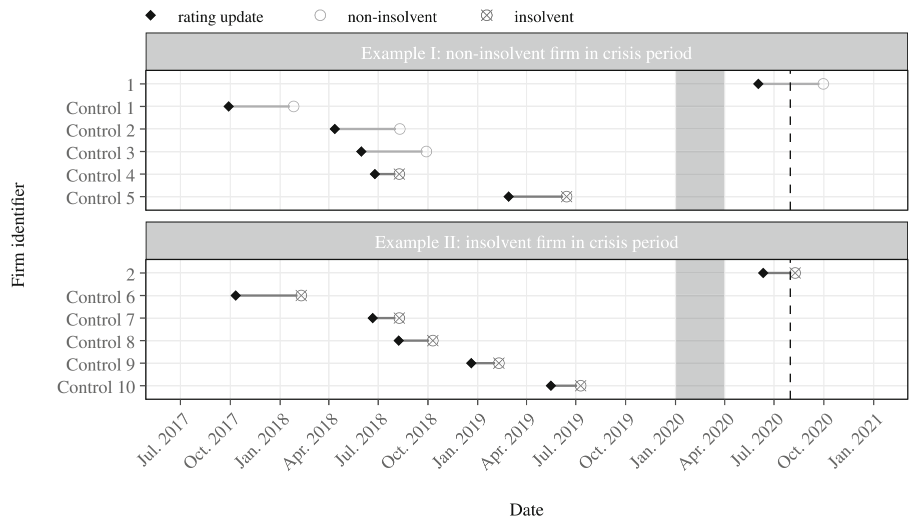


Fig. 2 Matching: illustration. Note: The figure illustrates the nearest neighbor matching for two micro-enterprises in the accommodation and catering sector. In the top panel, we see that firm 1 experienced a rating update in the crisis period which did not result in an insolvency filing. Furthermore, we see, however, that two out of the $k = 5$ nearest neighbors from the pre-crisis control period filed for insolvency after they received a very similar rating update. This signals that firm 1, given its financial information, faces a relatively high insolvency risk as almost half of its nearest neighbors indeed went bankrupt in times without policy intervention. The bottom panel shows the same approach but for firm 2 which filed for insolvency shortly after its rating update during the crisis period. We see that all of the nearest neighbors also filed for insolvency and thus closely reflect the actual survival status of firm 2. If we do not observe an insolvency filing four months after the rating update, we treat the update as non-insolvent. Therefore, the time between rating update and the non-insolvent labelling in the visualization always spans 4 months. The area shaded in gray highlights a transitional phase which we intentionally exclude from our analysis since assignment of observations falling in that phase to either the pre-crisis or the crisis period is not straightforward. The dashed vertical line at the end of July 2020 signals that we only have credit rating updates available up to this point. Note, however, that we observe insolvency filings beyond this point in time

COVID-19 period. In the following section, we introduce all of these matching variables and provide some descriptive statistics.

In our data used for the estimation of the insolvency gap firm survival status, f_{t+4}, serves as outcome variable. It is equal to 1 if the company has filed for insolvency no more than four months after its last rating update. If the firm has not gone bankrupt or it has filed insolvency more than four months after its latest rating update, it is 0. This means that we take four months as maximum time lag between a credit rating update and the date at which the respective firm has filed its bankruptcy to count the rating update as being predictive for the subsequent insolvency filing. We choose this threshold for two reasons. First, we want to ensure that the rating update has a high information content in predicting a potential insolvency filing. If the date of bankruptcy lies more than 4 months after the credit update, it is likely that the update does not reflect the reasons why the company went bankrupt. A more recent update of the firm's rating (if that existed) would be necessary to capture the company's financial deterioration that contributed to the subsequent insolvency. Second, the COVID-19 period for which we have information on credit rating updates spans 4 months from April 2020 to the end of July 2020. Thus, for the latest in-crisis rating updates in July 2020, we can observe the firm's survival status at most 4 months until November 2020 (the time of writing this paper). Therefore, the maximum forecasting horizon for the rating updates observed in the crisis period is limited to 4 months.

The most important variable in finding counterfactual survival states in the matching procedure is Creditreform's credit rating index since it is the basis for the calculation of the credit rating updates. The credit rating is calculated by Creditreform on the basis of a rich information set relevant to assess a company's creditworthiness. The metrics considered in calculating the rating include, among other things, information on the firm's payment discipline, its legal form, credit evaluations of banks, credit line limits and risk indicators based on the firm's financial accounts (if applicable) (Creditreform, 2020b). Creditreform attaches different weights to these metrics according to their relevance on determining a firm's risk of credit default and calculates an overall credit rating score which ranges from 100 to 500.[8] The higher the score, the worse the firm's creditworthiness and thus the higher the risk of insolvency. In fact, Creditreform's rating index has a high forecasting quality to assess a firm's credit default risk (Creditreform, 2020b). Assuming that a high credit default risk signals financial distress, which often results in insolvency, we use Creditreform's rating as the basis for predicting corporate insolvency risk. The prediction of corporate bankruptcy via a scoring model goes back to the seminal work of Altman (1968) and his development of the Z-score model. Similar to Creditreform's credit rating index, the Z-score model relies on several accounting-based indicators which are weighted and summed to obtain an overall score. This score then forms the basis for classifying companies as insolvent or non-insolvent (Altman, 2013). Today, this model approach is still used by many practitioners to predict firm insolvencies (Agarwal and Taffler, 2008).

Based on the credit rating index, we construct the following variables. Our main predictor variable is the *update* in the rating index, Δr_t, which is defined as the difference of the new rating assigned by Creditreform an the rating before the update ($\Delta r_t = r_t - r_{t-x}$).[9] Given the logic of the rating index, a positive sign in the rating update reflects a downgrade in financial solvency, a negative sign reflects an improvement in the rating, i.e. an upgrade of the company's financial standing. The amount of the down-/upgrade reflects how severely the company's financial standing has changed.[10]

Apart from the rating update, we also consider the rating before the upgrade, r_{t-x}, as a matching variable when predicting counterfactual insolvency states. This allows us to control for the location of the company in the rating distribution and consequently how high the default risk was before the down-/upgrade. Moreover, we form two additional variables from the firm's credit rating information, both of which control for the medium-term path of the firm's financial standing. First, we count the number of downgrades in the three years preceding the update at hand, d_t. Second, we calculate the average credit rating in the three years prior to the current update under consideration, $\bar{r}_t$.[11] Finally, we consider firm age, a_t, as further matching variable acknowledging that younger firms tend to be more prone to insolvency.

Table 3 shows descriptive statistics of the variables considered in the matching procedure. We see that an update which is followed by a bankruptcy filing relates to a downgrade of close to 70 scoring points on average. This is a substantial deterioration in the rating index compared to an update which is not followed by an insolvency filing. In fact, the difference in means between non-insolvency-related updates and insolvency-related updates, as reported in column 'Δ Mean', amounts to more than 65 index points and is statistically significant. For all other matching variables, we also find statistically meaningful differences suggesting that firms which go bankrupt have a worse credit rating both short-term and mid-term, have experienced more downgrades in the past and are younger on average. The economically and statistically significant differences between non-insolvency-related and insolvency-related credit updates across all variables

[8] The credit rating index suffers from a discontinuity as in case of a 'insufficient' creditworthiness it takes on a value of 600 (Creditreform, 2020a). We truncate credit ratings of 600 to a value 500—the worst possible rating in our analysis. We do so since our main predictor variable is the *update* in the rating index which can only be reasonably calculated if the index has continuous support.

[9] Reassessments of the rating is conducted in an irregular fashion such that the time between two updates, x, varies. On average, the time between two updates equals 20 months.

[10] Note that we define a rating update as a reassessment of the company's creditworthiness performed by Creditreform. We have precise information on the date of reassessment, which allows us to accurately assign the update to either the pre-crisis or the crisis period and also to accurately match the updates with insolvency dates. It should also be noted that a reassessment does not necessarily lead to a change in the rating index. If the creditworthiness of the company has not changed since the last rating, the company gets assigned the same index as before, resulting in a value of 0 for Δr_t.

[11] For example, for a credit rating observation in July 2017, we count how often the firm experienced a downgrade over the period June 2014 to June 2017 and also calculate the average rating over that period.

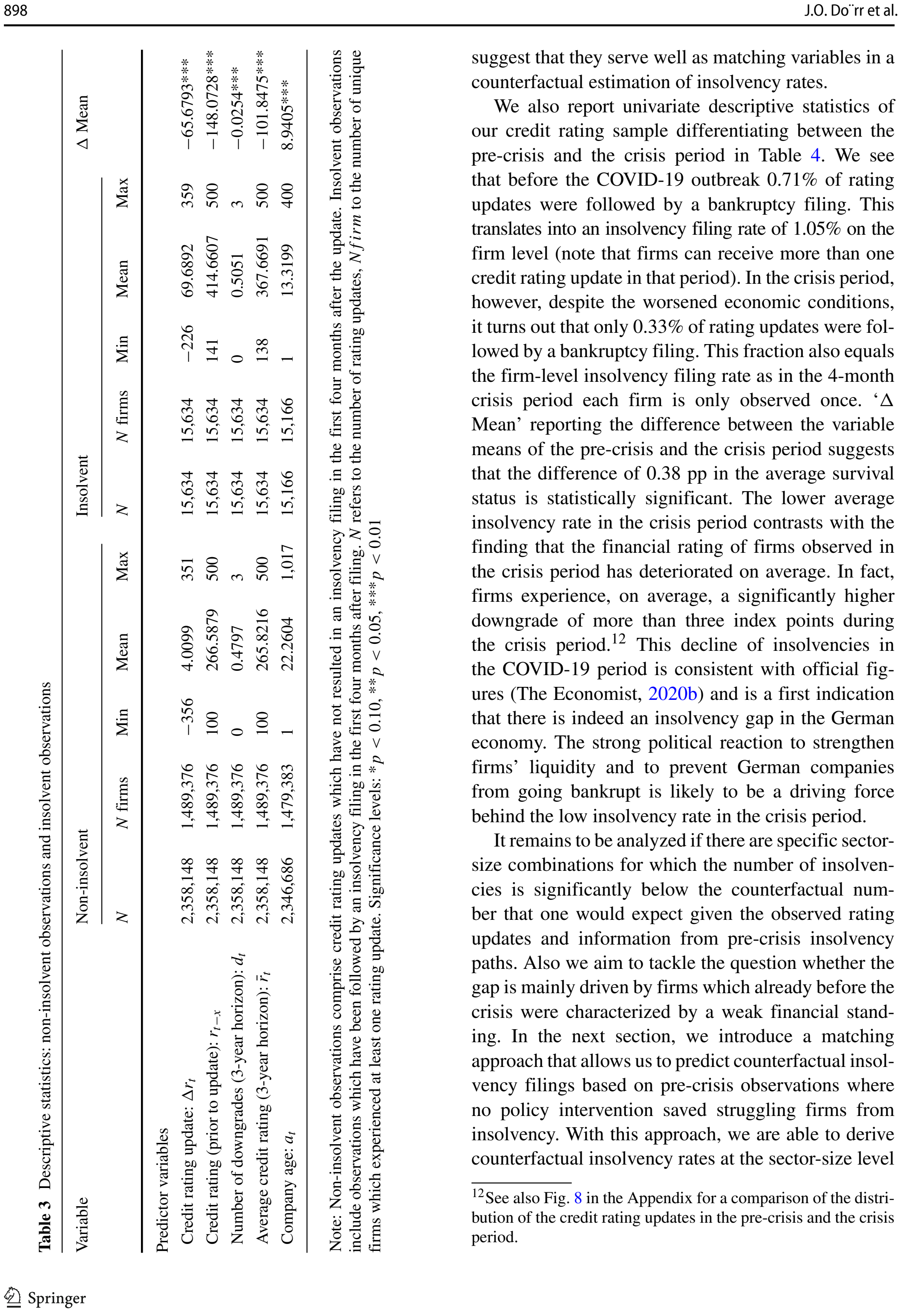


Table 3 Descriptive statistics: non-insolvent observations and insolvent observations

Variable	Non-insolvent					Insolvent					Δ Mean
	N	*N* firms	Min	Mean	Max	*N*	*N* firms	Min	Mean	Max	
Predictor variables											
Credit rating update: Δr_t	2,358,148	1,489,376	−356	4.0099	351	15,634	15,634	−226	69.6892	359	−65.6793***
Credit rating (prior to update): r_{t-x}	2,358,148	1,489,376	100	266.5879	500	15,634	15,634	141	414.6607	500	−148.0728***
Number of downgrades (3-year horizon): d_t	2,358,148	1,489,376	0	0.4797	3	15,634	15,634	0	0.5051	3	−0.0254***
Average credit rating (3-year horizon): $\bar{r}_t$	2,358,148	1,489,376	100	265.8216	500	15,634	15,634	138	367.6691	500	−101.8475***
Company age: a_t	2,346,686	1,479,383	1	22.2604	1,017	15,166	15,166	1	13.3199	400	8.9405***

Note: Non-insolvent observations comprise credit rating updates which have not resulted in an insolvency filing in the first four months after the update. Insolvent observations include observations which have been followed by an insolvency filing in the first four months after filing. N refers to the number of rating updates, $Nfirm$ to the number of unique firms which experienced at least one rating update. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

suggest that they serve well as matching variables in a counterfactual estimation of insolvency rates.

We also report univariate descriptive statistics of our credit rating sample differentiating between the pre-crisis and the crisis period in Table 4. We see that before the COVID-19 outbreak 0.71% of rating updates were followed by a bankruptcy filing. This translates into an insolvency filing rate of 1.05% on the firm level (note that firms can receive more than one credit rating update in that period). In the crisis period, however, despite the worsened economic conditions, it turns out that only 0.33% of rating updates were followed by a bankruptcy filing. This fraction also equals the firm-level insolvency filing rate as in the 4-month crisis period each firm is only observed once. 'Δ Mean' reporting the difference between the variable means of the pre-crisis and the crisis period suggests that the difference of 0.38 pp in the average survival status is statistically significant. The lower average insolvency rate in the crisis period contrasts with the finding that the financial rating of firms observed in the crisis period has deteriorated on average. In fact, firms experience, on average, a significantly higher downgrade of more than three index points during the crisis period.[12] This decline of insolvencies in the COVID-19 period is consistent with official figures (The Economist, 2020b) and is a first indication that there is indeed an insolvency gap in the German economy. The strong political reaction to strengthen firms' liquidity and to prevent German companies from going bankrupt is likely to be a driving force behind the low insolvency rate in the crisis period.

It remains to be analyzed if there are specific sector-size combinations for which the number of insolvencies is significantly below the counterfactual number that one would expect given the observed rating updates and information from pre-crisis insolvency paths. Also we aim to tackle the question whether the gap is mainly driven by firms which already before the crisis were characterized by a weak financial standing. In the next section, we introduce a matching approach that allows us to predict counterfactual insolvency filings based on pre-crisis observations where no policy intervention saved struggling firms from insolvency. With this approach, we are able to derive counterfactual insolvency rates at the sector-size level

[12] See also Fig. 8 in the Appendix for a comparison of the distribution of the credit rating updates in the pre-crisis and the crisis period.

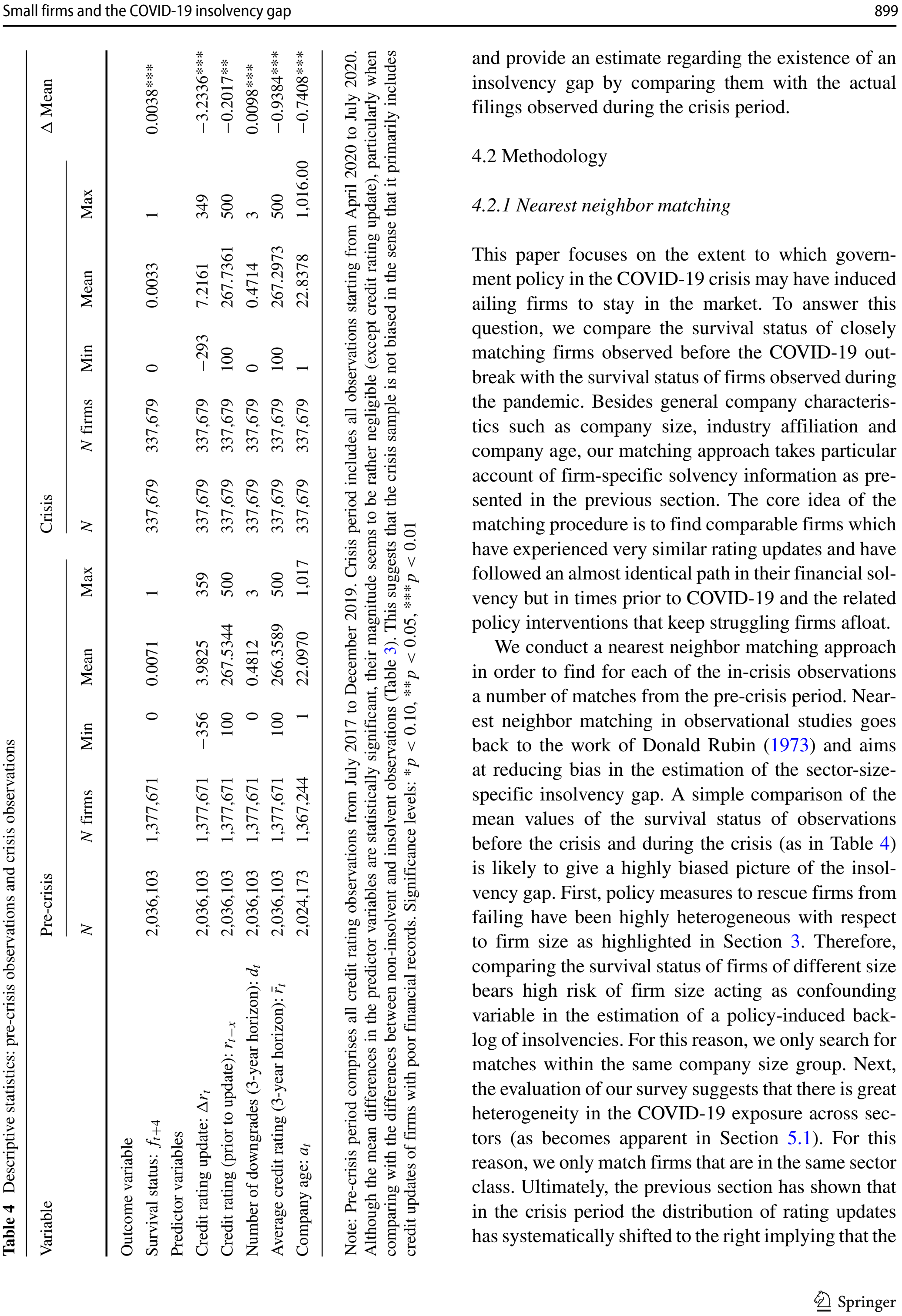


Table 4 Descriptive statistics: pre-crisis observations and crisis observations

Variable	Pre-crisis					Crisis					Δ Mean
	N	*N* firms	Min	Mean	Max	*N*	*N* firms	Min	Mean	Max	
Outcome variable											
Survival status: f_{t+4}	2,036,103	1,377,671	0	0.0071	1	337,679	337,679	0	0.0033	1	0.0038***
Predictor variables											
Credit rating update: Δr_t	2,036,103	1,377,671	−356	3.9825	359	337,679	337,679	−293	7.2161	349	−3.2336***
Credit rating (prior to update): r_{t-x}	2,036,103	1,377,671	100	267.5344	500	337,679	337,679	100	267.7361	500	−0.2017**
Number of downgrades (3-year horizon): d_t	2,036,103	1,377,671	0	0.4812	3	337,679	337,679	0	0.4714	3	0.0098***
Average credit rating (3-year horizon): $\bar{r}_t$	2,036,103	1,377,671	100	266.3589	500	337,679	337,679	100	267.2973	500	−0.9384***
Company age: a_t	2,024,173	1,367,244	1	22.0970	1,017	337,679	337,679	1	22.8378	1,016.00	−0.7408***

Note: Pre-crisis period comprises all credit rating observations from July 2017 to December 2019. Crisis period includes all observations starting from April 2020 to July 2020. Although the mean differences in the predictor variables are statistically significant, their magnitude seems to be rather negligible (except credit rating update), particularly when comparing with the differences between non-insolvent and insolvent observations (Table 3). This suggests that the crisis sample is not biased in the sense that it primarily includes credit updates of firms with poor financial records. Significance levels: *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

and provide an estimate regarding the existence of an insolvency gap by comparing them with the actual filings observed during the crisis period.

4.2 Methodology

4.2.1 Nearest neighbor matching

This paper focuses on the extent to which government policy in the COVID-19 crisis may have induced ailing firms to stay in the market. To answer this question, we compare the survival status of closely matching firms observed before the COVID-19 outbreak with the survival status of firms observed during the pandemic. Besides general company characteristics such as company size, industry affiliation and company age, our matching approach takes particular account of firm-specific solvency information as presented in the previous section. The core idea of the matching procedure is to find comparable firms which have experienced very similar rating updates and have followed an almost identical path in their financial solvency but in times prior to COVID-19 and the related policy interventions that keep struggling firms afloat.

We conduct a nearest neighbor matching approach in order to find for each of the in-crisis observations a number of matches from the pre-crisis period. Nearest neighbor matching in observational studies goes back to the work of Donald Rubin (1973) and aims at reducing bias in the estimation of the sector-size-specific insolvency gap. A simple comparison of the mean values of the survival status of observations before the crisis and during the crisis (as in Table 4) is likely to give a highly biased picture of the insolvency gap. First, policy measures to rescue firms from failing have been highly heterogeneous with respect to firm size as highlighted in Section 3. Therefore, comparing the survival status of firms of different size bears high risk of firm size acting as confounding variable in the estimation of a policy-induced backlog of insolvencies. For this reason, we only search for matches within the same company size group. Next, the evaluation of our survey suggests that there is great heterogeneity in the COVID-19 exposure across sectors (as becomes apparent in Section 5.1). For this reason, we only match firms that are in the same sector class. Ultimately, the previous section has shown that in the crisis period the distribution of rating updates has systematically shifted to the right implying that the in-crisis observations have, on average, experienced larger downgrades in their ratings. For an unbiased estimation of the insolvency gap, this shift needs to be controlled for. Our nearest neighbor matching aligns the in-crisis distribution of updates with the distribution of matched observations as we put a strict caliper on the credit rating variable when searching for matching observations. In fact, comparing the distribution of the predictor variables between pre-crisis and crisis period before and after matching indicates that control observations and crisis observations are much more balanced after matching (see Table 12 in the Appendix for an assessment of covariate balance).

The details of our matching algorithm look as follows. Acknowledging the heterogeneity with respect to firm size and sector affiliation, we estimate the insolvency gap within each of the 52 sector-size combinations. Therefore, we only consider pre-crisis observations that share the same sector-size stratum as the crisis observation of interest. In that sense we perform exact matching on both sector affiliation and company size group. Next, within each sector-size stratum the algorithm selects for each in-crisis observation i the k nearest neighbors from the pre-crisis period which have the smallest distance from i. The maximum number of nearest neighbors, k, reflects the ratio of pre-crisis and crisis observations within each sector-size stratum. Distance is measured by the Mahalanobis distance metric (Rubin, 1980), MD, which is computed on all predictor variables $\boldsymbol{X} = (\Delta r_t \ r_{t-x} \ d_t \ \bar{r}_t \ a_t)'$. For the key predictor variable, Δr_t, we additionally impose a caliper, c, of 0.25 standard deviations. Thus, a pre-crisis observation, j, only falls under the k nearest neighbors if it does not exceed the caliper on Δr_t.

$$MD_{ij}=\begin{cases}(\boldsymbol{X}_i-\boldsymbol{X}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j) & \text{if } |\Delta r_{t,i}-\Delta r_{t,j}|\leq c\\ \infty & \text{if } |\Delta r_{t,i}-\Delta r_{t,j}|>c\end{cases}$$

with $\boldsymbol{\Sigma}$ as the variance covariance matrix of $\boldsymbol{X}$ in the pooled sample of in-crisis and all pre-crisis observations. The strict caliper implies that the number of matches on each crisis observation can be smaller than k or, in case that there is no control observation fulfilling the caliper condition, there may even be no match. If this the case, the crisis observation for which no match could be found is disregarded from further analysis. Moreover, we conduct matching with replacement allowing pre-crisis units to match to more than one crisis observation. In the outcome analysis, this requires us to consider weights which reflect whether a pre-crisis unit falls in the matched sample more than once. In Section 5.2 where we estimate the insolvency gap on the sector-size level, we need to consider these weights for inference (Stuart, 2010). In this way, we can not only predict the crisis observations' probability to file for bankruptcy if there was no policy intervention but also make a statement whether the differences between the observed insolvency rates and the predicted counterfactual insolvency rates on the sector-size level are statistically significant.

Before presenting the results of the counterfactual insolvency rate prediction and insolvency gap estimation, we use our survey results in the next section to show how the pandemic affected sectors to varying degrees. The observed heterogeneity in sector exposure motivates our further empirical analysis.

5 Empirical results

5.1 COVID-19 exposure and firm characteristics

Anecdotal evidence suggests that industries are asymmetrically affected by the COVID-19 recession because lockdown measures as well as supply and demand effects differed between sectors. To verify this observation, we empirically investigate to what extent the economic effects of the COVID-19 crisis have asymmetrically hit sectors by making use of our survey data. In addition, we analyze whether firm size and the pre-crisis credit rating is correlated with the perceived shock by the COVID-19 recession at the firm level.

The regression results of the analyses are shown in Table 5. Model (1) reveals that the COVID-19 Exposure Index indeed significantly differs between sectors. We choose chemicals and pharmaceuticals as reference category since this sector is least negatively affected. The sectors accommodation and catering as well as creative industry and entertainment experience very strong and significant negative shocks in comparison to the baseline sector. This is in line with the strong restrictions experienced in these sectors. Since the business activities in these sectors often require direct human interactions, corresponding companies have been severely affected by lockdown measures. Interestingly, firm size categories show no statistically

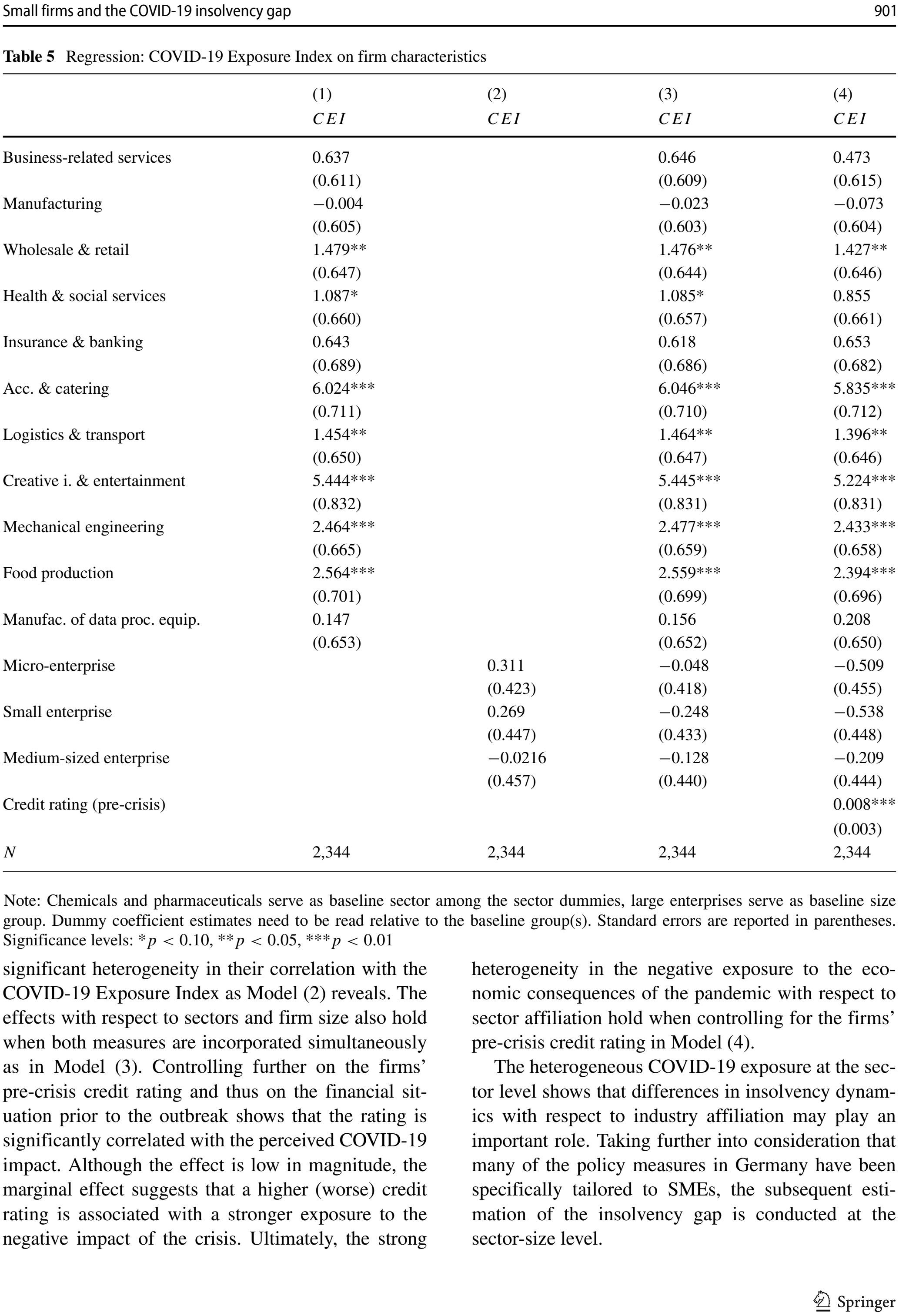


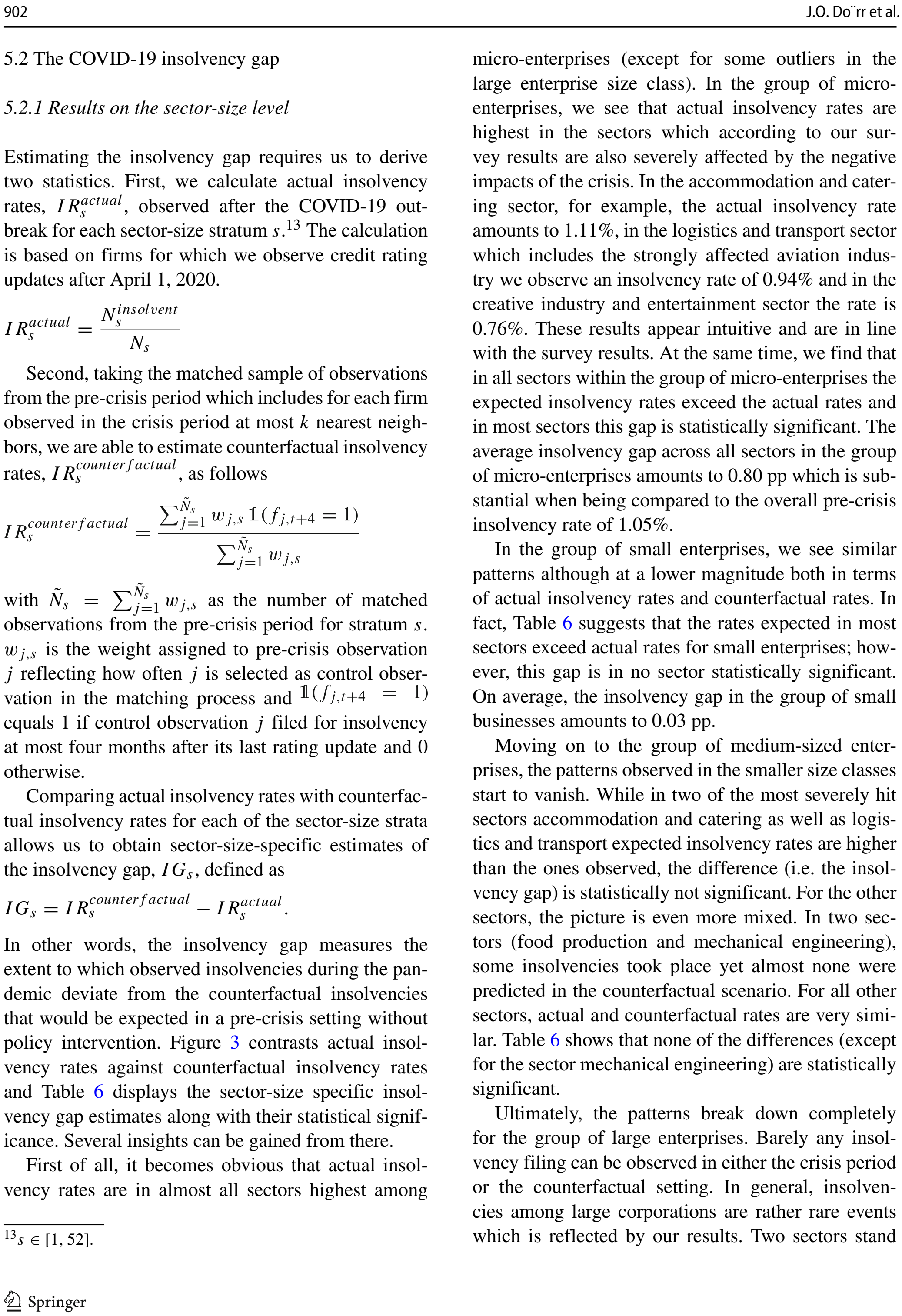
Table 5 Regression: COVID-19 Exposure Index on firm characteristics

	(1) CEI	(2) CEI	(3) CEI	(4) CEI
Business-related services	0.637		0.646	0.473
	(0.611)		(0.609)	(0.615)
Manufacturing	−0.004		−0.023	−0.073
	(0.605)		(0.603)	(0.604)
Wholesale & retail	1.479**		1.476**	1.427**
	(0.647)		(0.644)	(0.646)
Health & social services	1.087*		1.085*	0.855
	(0.660)		(0.657)	(0.661)
Insurance & banking	0.643		0.618	0.653
	(0.689)		(0.686)	(0.682)
Acc. & catering	6.024***		6.046***	5.835***
	(0.711)		(0.710)	(0.712)
Logistics & transport	1.454**		1.464**	1.396**
	(0.650)		(0.647)	(0.646)
Creative i. & entertainment	5.444***		5.445***	5.224***
	(0.832)		(0.831)	(0.831)
Mechanical engineering	2.464***		2.477***	2.433***
	(0.665)		(0.659)	(0.658)
Food production	2.564***		2.559***	2.394***
	(0.701)		(0.699)	(0.696)
Manufac. of data proc. equip.	0.147		0.156	0.208
	(0.653)		(0.652)	(0.650)
Micro-enterprise		0.311	−0.048	−0.509
		(0.423)	(0.418)	(0.455)
Small enterprise		0.269	−0.248	−0.538
		(0.447)	(0.433)	(0.448)
Medium-sized enterprise		−0.0216	−0.128	−0.209
		(0.457)	(0.440)	(0.444)
Credit rating (pre-crisis)				0.008***
				(0.003)
N	2,344	2,344	2,344	2,344

Note: Chemicals and pharmaceuticals serve as baseline sector among the sector dummies, large enterprises serve as baseline size group. Dummy coefficient estimates need to be read relative to the baseline group(s). Standard errors are reported in parentheses. Significance levels: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$

significant heterogeneity in their correlation with the COVID-19 Exposure Index as Model (2) reveals. The effects with respect to sectors and firm size also hold when both measures are incorporated simultaneously as in Model (3). Controlling further on the firms' pre-crisis credit rating and thus on the financial situation prior to the outbreak shows that the rating is significantly correlated with the perceived COVID-19 impact. Although the effect is low in magnitude, the marginal effect suggests that a higher (worse) credit rating is associated with a stronger exposure to the negative impact of the crisis. Ultimately, the strong heterogeneity in the negative exposure to the economic consequences of the pandemic with respect to sector affiliation hold when controlling for the firms' pre-crisis credit rating in Model (4).

The heterogeneous COVID-19 exposure at the sector level shows that differences in insolvency dynamics with respect to industry affiliation may play an important role. Taking further into consideration that many of the policy measures in Germany have been specifically tailored to SMEs, the subsequent estimation of the insolvency gap is conducted at the sector-size level.



5.2 The COVID-19 insolvency gap

5.2.1 Results on the sector-size level

Estimating the insolvency gap requires us to derive two statistics. First, we calculate actual insolvency rates, IR_s^{actual}, observed after the COVID-19 outbreak for each sector-size stratum s.[13] The calculation is based on firms for which we observe credit rating updates after April 1, 2020.

$$IR_s^{actual} = \frac{N_s^{insolvent}}{N_s}$$

Second, taking the matched sample of observations from the pre-crisis period which includes for each firm observed in the crisis period at most k nearest neighbors, we are able to estimate counterfactual insolvency rates, $IR_s^{counterfactual}$, as follows

$$IR_s^{counterfactual} = \frac{\sum_{j=1}^{\tilde{N}_s} w_{j,s}\, \mathbb{1}(f_{j,t+4} = 1)}{\sum_{j=1}^{\tilde{N}_s} w_{j,s}}$$

with $\tilde{N}_s = \sum_{j=1}^{\tilde{N}_s} w_{j,s}$ as the number of matched observations from the pre-crisis period for stratum s. $w_{j,s}$ is the weight assigned to pre-crisis observation j reflecting how often j is selected as control observation in the matching process and $\mathbb{1}(f_{j,t+4} = 1)$ equals 1 if control observation j filed for insolvency at most four months after its last rating update and 0 otherwise.

Comparing actual insolvency rates with counterfactual insolvency rates for each of the sector-size strata allows us to obtain sector-size-specific estimates of the insolvency gap, IG_s, defined as

$$IG_s = IR_s^{counterfactual} - IR_s^{actual}.$$

In other words, the insolvency gap measures the extent to which observed insolvencies during the pandemic deviate from the counterfactual insolvencies that would be expected in a pre-crisis setting without policy intervention. Figure 3 contrasts actual insolvency rates against counterfactual insolvency rates and Table 6 displays the sector-size specific insolvency gap estimates along with their statistical significance. Several insights can be gained from there.

First of all, it becomes obvious that actual insolvency rates are in almost all sectors highest among micro-enterprises (except for some outliers in the large enterprise size class). In the group of micro-enterprises, we see that actual insolvency rates are highest in the sectors which according to our survey results are also severely affected by the negative impacts of the crisis. In the accommodation and catering sector, for example, the actual insolvency rate amounts to 1.11%, in the logistics and transport sector which includes the strongly affected aviation industry we observe an insolvency rate of 0.94% and in the creative industry and entertainment sector the rate is 0.76%. These results appear intuitive and are in line with the survey results. At the same time, we find that in all sectors within the group of micro-enterprises the expected insolvency rates exceed the actual rates and in most sectors this gap is statistically significant. The average insolvency gap across all sectors in the group of micro-enterprises amounts to 0.80 pp which is substantial when being compared to the overall pre-crisis insolvency rate of 1.05%.

In the group of small enterprises, we see similar patterns although at a lower magnitude both in terms of actual insolvency rates and counterfactual rates. In fact, Table 6 suggests that the rates expected in most sectors exceed actual rates for small enterprises; however, this gap is in no sector statistically significant. On average, the insolvency gap in the group of small businesses amounts to 0.03 pp.

Moving on to the group of medium-sized enterprises, the patterns observed in the smaller size classes start to vanish. While in two of the most severely hit sectors accommodation and catering as well as logistics and transport expected insolvency rates are higher than the ones observed, the difference (i.e. the insolvency gap) is statistically not significant. For the other sectors, the picture is even more mixed. In two sectors (food production and mechanical engineering), some insolvencies took place yet almost none were predicted in the counterfactual scenario. For all other sectors, actual and counterfactual rates are very similar. Table 6 shows that none of the differences (except for the sector mechanical engineering) are statistically significant.

Ultimately, the patterns break down completely for the group of large enterprises. Barely any insolvency filing can be observed in either the crisis period or the counterfactual setting. In general, insolvencies among large corporations are rather rare events which is reflected by our results. Two sectors stand

[13] $s \in [1, 52]$.

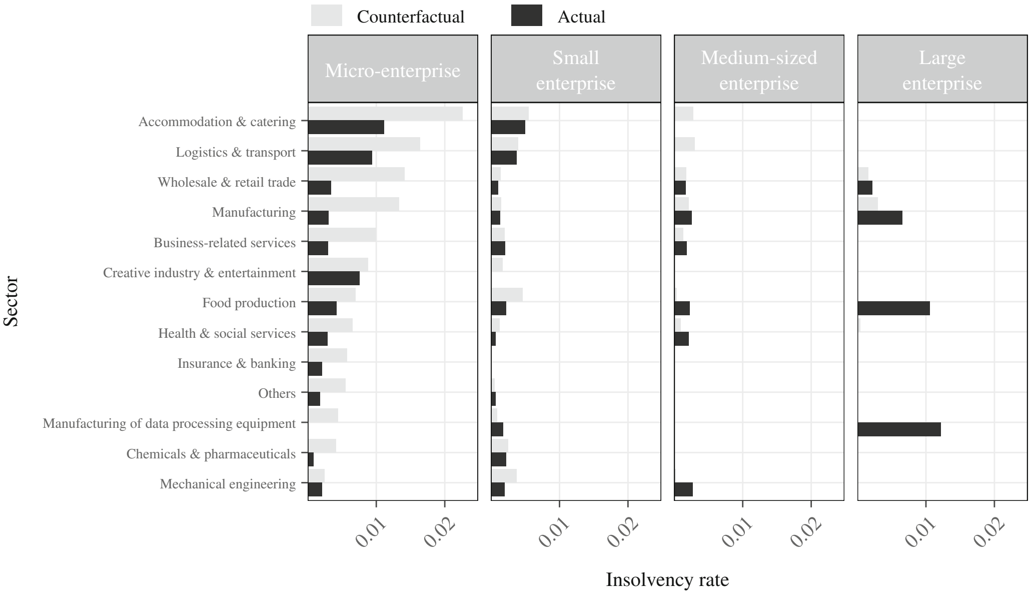


Fig. 3 Actual and counterfactual insolvency rates. Note: The figure displays actual insolvency rates with estimated counterfactual insolvency rates for each of the $S = 52$ sector-size strata

Table 6 Outcome analysis: insolvency gap estimation results

Sector affiliation	Size of company			
	Micro	Small	Medium	Large
	IG_s	IG_s	IG_s	IG_s
Accommodation & catering	+0.0115***	+0.0005	+0.0028	0.0000
Logistics & transport	+0.0070***	+0.0002	+0.0030	0.0000
Wholesale & retail trade	+0.0107***	+0.0004	+0.0001	−0.0006
Manufacturing	+0.0103***	+0.0002	−0.0004	−0.0035
Business-related services	+0.0070***	−0.0001	−0.0005	0.0000
Creative industry & entertainment	+0.0012	+0.0017	0.0000	0.0000
Food production	+0.0027	+0.0024	−0.0019	−0.0105**
Health & social services	+0.0037***	+0.0005	−0.0011	+0.0004
Insurance & banking	+0.0037***	0.0000	0.0000	0.0000
Others	+0.0037***	−0.0002	0.0000	0.0000
Manufacturing of data processing equipment	+0.0044*	−0.0009	0.0000	−0.0122*
Chemicals & pharmaceuticals	+0.0033*	+0.0003	0.0000	0.0000
Mechanical engineering	+0.0003	+0.0018	−0.0025***	0.0000

Note: Significance levels: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$. Statistical significance is based on the χ^2-test for equality in the insolvency proportions in the actual and counterfactual samples using Rao-Scott corrections to the χ^2 statistic (Rao & Scott, 1981) to account for the matching weights out with high actual insolvency rates: food production and manufacturing of data processing equipment. Both cases are somewhat special as they are driven by only one insolvency for which no insolvent pre-crisis control observation with comparable financial characteristics exists. Thus, one needs to be cautious when interpreting the results of the large size class.

The finding that counterfactual insolvency rates persistently, and in most sectors also significantly, exceed actual rates among micro-enterprises strongly suggests that there is a substantial backlog of insolvencies in this size class. As company size increases, the backlog of insolvencies gradually vanishes which is in line with our hypothesis that Germany's fiscal policy response in the COVID-19 crisis disproportionately favored the survival of smaller companies. Both the temporary change in Germany's insolvency regime and the high provision of liquidity subsidies allowed especially micro-enterprises to stay in the market. We argue that the temporary suspension of the obligation to file for insolvencies has made it particularly easy for smaller firms to use the amendment as a loophole to avert insolvency proceedings. Since disclosure requirements are more limited the smaller a company is, it becomes particularly difficult for policy makers to enforce insolvency filings among non-filing small firms. This becomes particularly problematic if the non-filing firm does not fulfill the criteria to be eligible for the suspension as it enables these companies to further absorb state subsidies. Similarly, the early on provision of direct and indirect liquidity without red tape has targeted smaller firms in particular and thus enabled them to bridge plummeting revenues in a situation in which they usually would have been forced out of the market due to illiquidity.

In order to better understand the magnitude of the insolvency gap, it is possible to aggregate and convert the insolvency gap estimates on the distinct sector-size levels into an absolute number describing the overall backlog of insolvencies (see Table 11 in the Appendix). Based on the total number of economically active companies in Germany, we estimate that the insolvency gap makes up around 25,000 companies as shown in Fig. 4. The figure reveals two further aspects. Firstly, the time series shows that during the last economic shock, the Great Recession of 2008–2009, the number of insolvencies noticeably increased, which in light of the Schumpeterian cleansing mechanism is an expected response in business dynamics. Secondly, in contrast to the Great Recession, it can be seen that in the current crisis the actual number of corporate insolvencies has declined. The

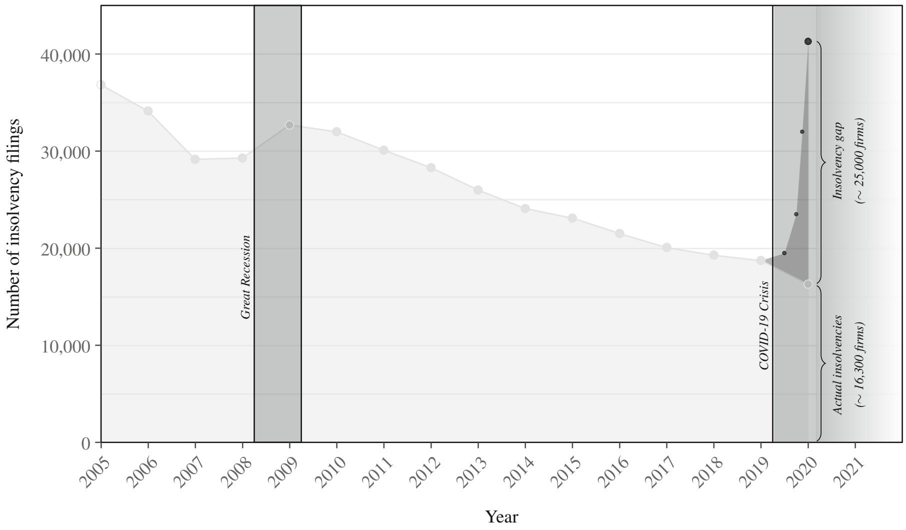


Fig. 4 Evolution of corporate insolvencies in Germany and the COVID-19 insolvency gap. Note: The figure displays yearly absolute number of corporate insolvencies in Germany according to official statistics of the Federal Statistical Office. In 2020, both actual number of insolvencies as well as the evolution of insolvencies in the counterfactual scenario are displayed. Shaded areas reflect months of economic downturn

observation that bankruptcy filings are lower in an economic crisis than in non-crisis times underpins that the large-scale governmental support programs have led to substantial distortions in business dynamics. Indeed, policy measures in Germany have prevented a significant number of companies from insolvency. The crucial question is, *which* firms were saved from insolvency proceedings. The following section further narrows down this question by incorporating the firms' pre-crisis financial standing in the estimation of the insolvency gap.

5.2.2 The insolvency gap and firm viability

In order to examine whether the insolvency gap is driven by companies that are characterized by a poor financial standing before the crisis and had faced a relatively high risk of market exit when the pandemic hit,

Table 7 Outcome analysis: insolvency gap estimation results incorporating firms' pre-crisis financial condition

Viability	Sector affiliation	Size of company			
		Micro	Small	Medium	Large
		IG_s	IG_s	IG_s	IG_s
Strong financial standing (pre-crisis)					
	Accommodation & catering	−0.0029***	−0.0013	0.0000	0.0000
	Logistics & transport	−0.0008	−0.0001	+0.0003	0.0000
	Wholesale & retail trade	+0.0003	−0.0001	+0.0005	−0.0007
	Manufacturing	−0.0001	−0.0001	−0.0008	−0.0038
	Business-related services	−0.0002	−0.0006	−0.0010	0.0000
	Creative industry & entertainment	−0.0025**	0.0000	0.0000	0.0000
	Food production	−0.0007	+0.0020	−0.0027*	−0.0112*
	Health & social services	−0.0001	−0.0004	−0.0003	0.0000
	Insurance & banking	+0.0007	0.0000	0.0000	0.0000
	Others	−0.0002	+0.0002	0.0000	0.0000
	Manufacturing of data processing equipment	+0.0007	−0.0015	0.0000	−0.0127*
	Chemicals & pharmaceuticals	+0.0017	+0.0020	0.0000	0.0000
	Mechanical engineering	−0.0003	−0.0017	−0.0015*	0.0000
Weak financial standing (pre-crisis)					
	Accommodation & catering	+0.0171***	+0.0018	+0.0030	0.0000
	Logistics & transport	+0.0128***	+0.0004	+0.0049	0.0000
	Wholesale & retail trade	+0.0196***	+0.0020	−0.0035	0.0000
	Manufacturing	+0.0184***	+0.0015	+0.0060	−0.0060
	Business-related services	+0.0122***	+0.0013	+0.0033	0.0000
	Creative industry & entertainment	+0.0029	+0.0032	0.0000	–
	Food production	+0.0051	+0.0030	+0.0025	0.0000
	Health & social services	+0.0059***	+0.0022	+0.0010	+0.0060
	Insurance & banking	+0.0055***	0.0000	0.0000	0.0000
	Others	+0.0073***	−0.0012	0.0000	0.0000
	Manufacturing of data processing equipment	+0.0065	0.0000	0.0000	0.0000
	Chemicals & pharmaceuticals	+0.0050	−0.0050	0.0000	0.0000
	Mechanical engineering	+0.0014	+0.0126*	−0.0056	0.0000

Note: The upper panel displays insolvency gap estimates for firms with 'strong financial standing' comprising all firms whose three year average credit index prior to the crisis is better than the median rating index. The lower panel shows results for companies with a 'weak financial standing' including those with a rating worse than the median rating. For large firms in creative and entertainment sector with weak financial standing the insolvency gap could not have been calculated as no firm in this strata has been observed during the crisis period. Significance levels: $*p < 0.10$, $**p < 0.05$, $***p < 0.01$. Statistical significance is based on the χ^2-test for equality in the insolvency proportions in the actual and counterfactual samples using Rao-Scott corrections to the χ^2 statistic (Rao & Scott, 1981) to account for the matching weights we split the sector-size strata further according to the observations' pre-crisis financial standing. More precisely, we split each sector-size strata into two further sub-strata. The first sub-strata contains all observations whose three year average credit rating prior to the crisis is better than the overall median rating index. We refer to these as observations with 'strong financial standing', viable to survive the crisis based on their pre-crisis conditions. The other sub-strata comprises all observations worse than the overall pre-crisis median rating. Firms falling in such sub-strata are referred to as having a 'weak financial standing'. Given their pre-crisis financial circumstances, we expect them to be more vulnerable to default in the current crisis or even if the pandemic had not hit the economy.

Table 7 shows the insolvency gap estimates analogous to Table 6 with the additional distinction between strata comprising financially strong companies and financially weak ones. Several aspects become apparent from these results. First, we observe that among micro-enterprises with above median credit rating (top panel) there is in almost no sector a significant deviation between actual and counterfactual insolvency rate. There are, however, two exceptions. Both in the accommodation and catering sector and the creative and entertainment sector observed insolvencies significantly exceed expected insolvencies. We know from our survey that these two sectors are by far the most affected industries. Given the severe impairments in these industries, it seems plausible that companies which had been rated relatively well before the pandemic nevertheless file for insolvency more frequently than the counterfactual estimation would suggest. This means that, despite the cushioning effect provided by fiscal policy, micro-firms with a strong pre-crisis rating filed for bankruptcy more frequently than would have been possible to learn from the financial paths of similarly strong firms in pre-crisis times. Most important, however, is the finding that among micro-enterprises the insolvency gap as backlog of expected insolvencies is *not* driven by firms with a strong financial standing prior to the crisis. In contrast, it is driven by less viable companies with a rating worse than the median rating. This results from the insolvency gap estimates among micro-enterprises with weak financial standing (bottom panel). It becomes apparent that throughout all sectors the counterfactual insolvency rates exceed the actual rates indicating a backlog of insolvencies which in the majority of sectors is not

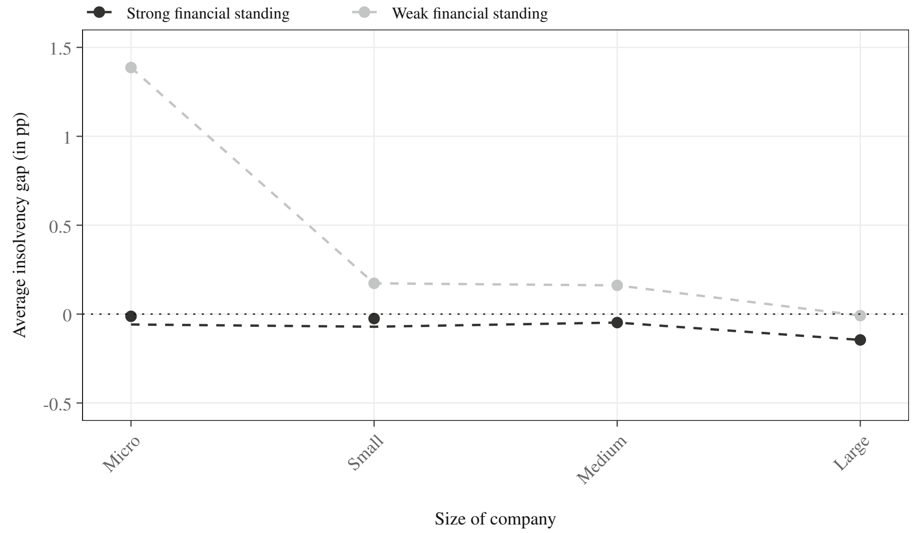


Fig. 5 Average insolvency rates by size class and pre-crisis viability. Note: The figure displays average insolvency gaps (weighted by the number of observations falling into the matched strata) distinguishing by company size and pre-crisis financial conditions. It becomes apparent that the backlog of insolvencies is strongly driven by micro-enterprises with weak credit rating prior to the crisis and also that with increasing firm size the gap becomes less pronounced among financially weak companies

only statistically significant but in some also substantial in magnitude. In fact, the gap amounts to more than 1.70 pp in accommodation and catering, wholesale and retail as well as manufacturing which is a substantial backlog when taking into consideration that the overall pre-crisis insolvency rate lies at 1.05%.

In the small business group, we see that in the strata with strong financial performance, the size and sign of the insolvency gap estimates are comparable to the results in the micro firm group (except for accommodation/catering and creative/entertainment sector). These results indicate that there is no significant gap in insolvency filings among small businesses with above median credit rating. In the strata of small and financially weak businesses, in turn, we observe for most sectors a positive sign in the insolvency gap estimation albeit only statistically significant in the mechanical engineering sector. Again, this suggests that also in the group of small firms the backlog of insolvencies is driven by companies with weak pre-crisis conditions even if magnitude and significance is less pronounced in comparison to the micro size group.

Similar to the results in Table 6 and in line with our hypothesis that the fiscal policy response in Germany disproportionately favored survival of smaller companies, the observed patterns for small and especially micro-sized firms gradually vanish with increasing firm size as shown in Fig. 5. Consequently, the insolvency gap estimates for medium-sized and large enterprises do not reveal clear patterns in the sign of the estimates nor significant deviations between observed and predicted rates (apart for some aforementioned exceptions).

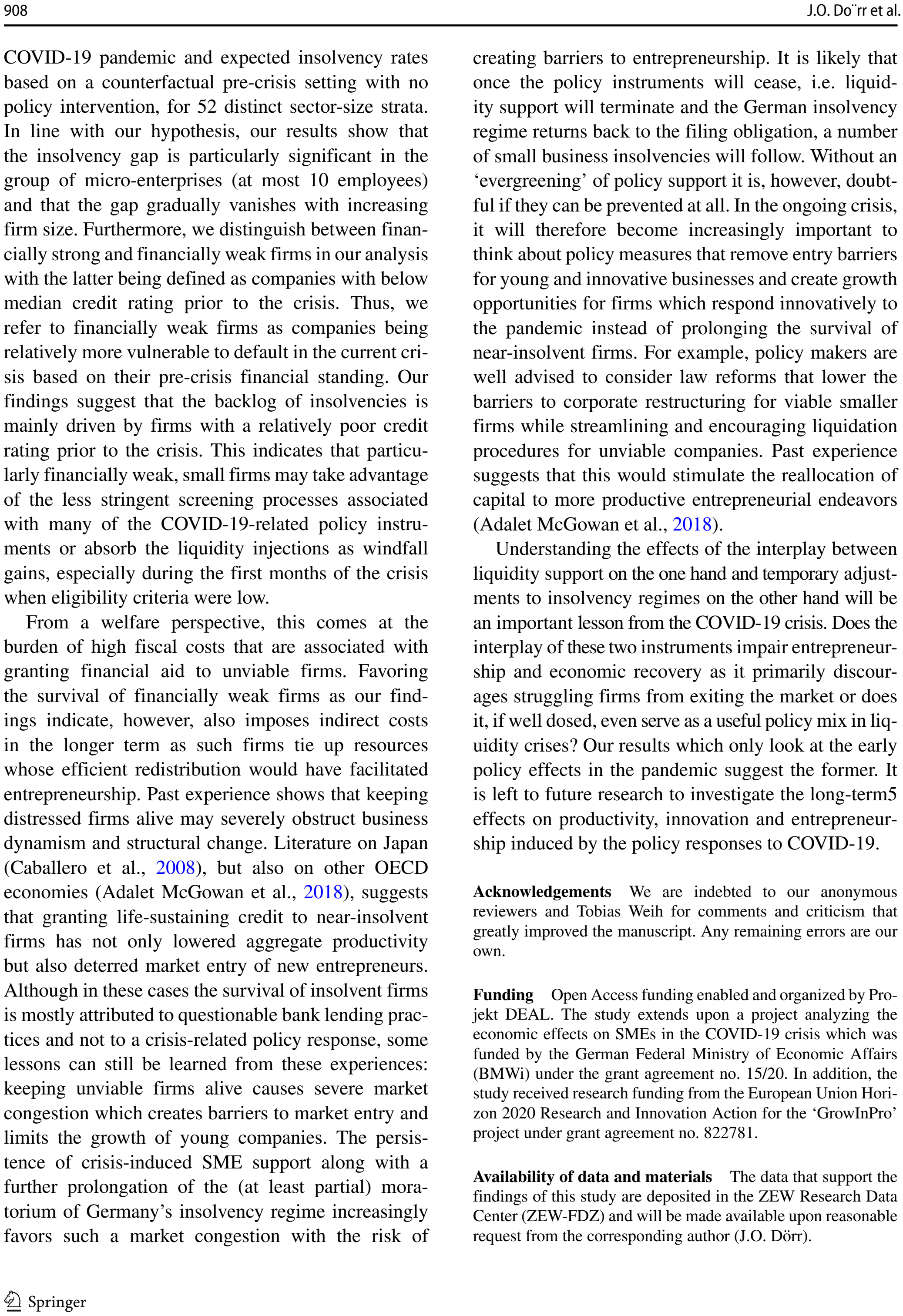
Our results show that the COVID-19-induced policy response has created a non-negligible insolvency gap that is strongly driven by micro-enterprises, which were already in a comparatively weak financial situation before the crisis.

This suggests that the early policy answer to dampen the economic impacts of the COVID-19 crisis has indeed hampered the natural cleansing effect typically observed during economic crises.

6 Discussion and conclusion

The ongoing COVID-19 crisis has placed a special role on policy in order to soften the adverse economic impacts faced by many firms. There is little doubt that, in the short term, liquidity subsidies and loan guarantees have been necessary to save companies under severe liquidity pressure from insolvency. In Germany, a country where fiscal policy played a crucial role in mitigating the crisis' impact, liquidity subsidies were accompanied by a temporary suspension of the insolvency regime. While both measures are different in design, they target the same objective: preventing an unprecedented wave of corporate insolvencies. Studying Germany's policy response, it becomes also apparent that a number of aid schemes were either explicitly designed to save smaller companies or at least implicitly favored the survival of particularly small entrepreneurial firms. This policy environment is the basis for our hypothesis that a substantial backlog of insolvencies has accumulated particularly among SMEs as a result of the COVID-19 policy response. If, however, support schemes postpone or even prevent the exit of financially weak SMEs, there is the danger of negative long-term effects on the entire economy. In fact, in the ongoing crisis it is likely that early liquidity issues increasingly translate into an erosion of firms' equity. Suspending bankruptcy proceedings of such over-indebted firms over a longer period of time not only is 'to deny reality' (The Economist 2020a, 3) but also hampers the efficient reallocation of resources. In this vein, economic crises also serve as cleansing mechanism to release resources from inefficient and non-innovative firms which typically find more productive use elsewhere. The early policy response of the German government not only has been targeted disproportionately at smaller firms but also did so with little screening mechanisms in place (see, for example, Federal Ministry for Economic Affairs and Energy 2020) rescuing companies from insolvency in a fairly indiscriminate manner.

Making use of both survey data and a unique and large dataset of firm-specific credit rating data along with information on firm insolvency filings, we investigate whether the German policy response has indeed caused distortions in the natural cleansing mechanism typically encountered in liquidity crises. While the policy response to the economic impact of COVID-19 in Germany suggests notable differences in firm size, our survey results reveal strong heterogeneity across economic sectors in their exposure to the adverse effects in the current crisis. With these findings, we estimate the extent of an insolvency gap, defined as the deviation of observed insolvency rates during the



COVID-19 pandemic and expected insolvency rates based on a counterfactual pre-crisis setting with no policy intervention, for 52 distinct sector-size strata. In line with our hypothesis, our results show that the insolvency gap is particularly significant in the group of micro-enterprises (at most 10 employees) and that the gap gradually vanishes with increasing firm size. Furthermore, we distinguish between financially strong and financially weak firms in our analysis with the latter being defined as companies with below median credit rating prior to the crisis. Thus, we refer to financially weak firms as companies being relatively more vulnerable to default in the current crisis based on their pre-crisis financial standing. Our findings suggest that the backlog of insolvencies is mainly driven by firms with a relatively poor credit rating prior to the crisis. This indicates that particularly financially weak, small firms may take advantage of the less stringent screening processes associated with many of the COVID-19-related policy instruments or absorb the liquidity injections as windfall gains, especially during the first months of the crisis when eligibility criteria were low.

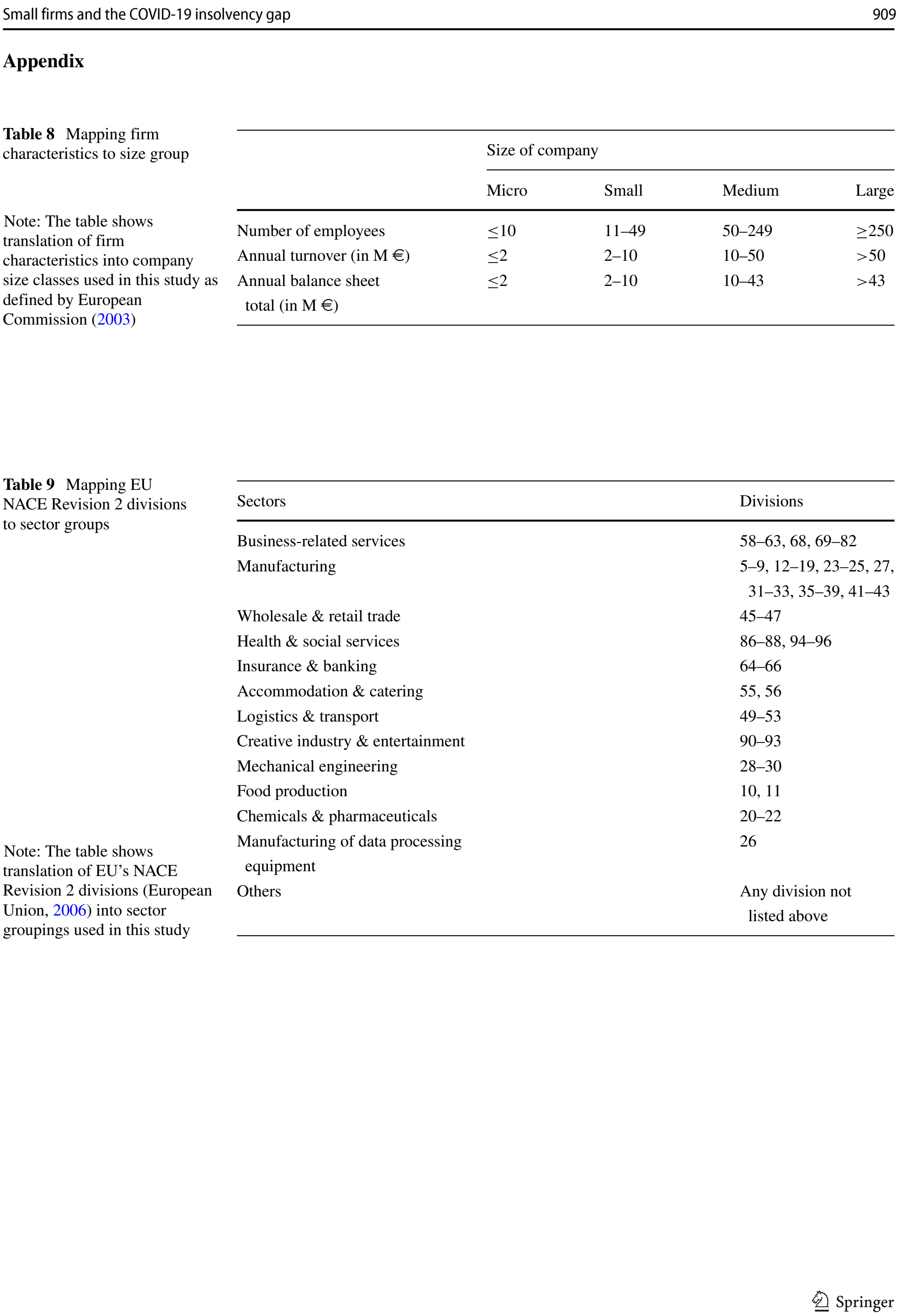
From a welfare perspective, this comes at the burden of high fiscal costs that are associated with granting financial aid to unviable firms. Favoring the survival of financially weak firms as our findings indicate, however, also imposes indirect costs in the longer term as such firms tie up resources whose efficient redistribution would have facilitated entrepreneurship. Past experience shows that keeping distressed firms alive may severely obstruct business dynamism and structural change. Literature on Japan (Caballero et al., 2008), but also on other OECD economies (Adalet McGowan et al., 2018), suggests that granting life-sustaining credit to near-insolvent firms has not only lowered aggregate productivity but also deterred market entry of new entrepreneurs. Although in these cases the survival of insolvent firms is mostly attributed to questionable bank lending practices and not to a crisis-related policy response, some lessons can still be learned from these experiences: keeping unviable firms alive causes severe market congestion which creates barriers to market entry and limits the growth of young companies. The persistence of crisis-induced SME support along with a further prolongation of the (at least partial) moratorium of Germany's insolvency regime increasingly favors such a market congestion with the risk of creating barriers to entrepreneurship. It is likely that once the policy instruments will cease, i.e. liquidity support will terminate and the German insolvency regime returns back to the filing obligation, a number of small business insolvencies will follow. Without an 'evergreening' of policy support it is, however, doubtful if they can be prevented at all. In the ongoing crisis, it will therefore become increasingly important to think about policy measures that remove entry barriers for young and innovative businesses and create growth opportunities for firms which respond innovatively to the pandemic instead of prolonging the survival of near-insolvent firms. For example, policy makers are well advised to consider law reforms that lower the barriers to corporate restructuring for viable smaller firms while streamlining and encouraging liquidation procedures for unviable companies. Past experience suggests that this would stimulate the reallocation of capital to more productive entrepreneurial endeavors (Adalet McGowan et al., 2018).

Understanding the effects of the interplay between liquidity support on the one hand and temporary adjustments to insolvency regimes on the other hand will be an important lesson from the COVID-19 crisis. Does the interplay of these two instruments impair entrepreneurship and economic recovery as it primarily discourages struggling firms from exiting the market or does it, if well dosed, even serve as a useful policy mix in liquidity crises? Our results which only look at the early policy effects in the pandemic suggest the former. It is left to future research to investigate the long-term5 effects on productivity, innovation and entrepreneurship induced by the policy responses to COVID-19.

Acknowledgements We are indebted to our anonymous reviewers and Tobias Weih for comments and criticism that greatly improved the manuscript. Any remaining errors are our own.

Funding Open Access funding enabled and organized by Projekt DEAL. The study extends upon a project analyzing the economic effects on SMEs in the COVID-19 crisis which was funded by the German Federal Ministry of Economic Affairs (BMWi) under the grant agreement no. 15/20. In addition, the study received research funding from the European Union Horizon 2020 Research and Innovation Action for the 'GrowInPro' project under grant agreement no. 822781.

Availability of data and materials The data that support the findings of this study are deposited in the ZEW Research Data Center (ZEW-FDZ) and will be made available upon reasonable request from the corresponding author (J.O. Dörr).



Appendix

Table 8 Mapping firm characteristics to size group

	Size of company			
	Micro	Small	Medium	Large
Number of employees	≤10	11–49	50–249	≥250
Annual turnover (in M €)	≤2	2–10	10–50	>50
Annual balance sheet total (in M €)	≤2	2–10	10–43	>43

Note: The table shows translation of firm characteristics into company size classes used in this study as defined by European Commission (2003)

Table 9 Mapping EU NACE Revision 2 divisions to sector groups

Sectors	Divisions
Business-related services	58–63, 68, 69–82
Manufacturing	5–9, 12–19, 23–25, 27, 31–33, 35–39, 41–43
Wholesale & retail trade	45–47
Health & social services	86–88, 94–96
Insurance & banking	64–66
Accommodation & catering	55, 56
Logistics & transport	49–53
Creative industry & entertainment	90–93
Mechanical engineering	28–30
Food production	10, 11
Chemicals & pharmaceuticals	20–22
Manufacturing of data processing equipment	26
Others	Any division not listed above

Note: The table shows translation of EU's NACE Revision 2 divisions (European Union, 2006) into sector groupings used in this study

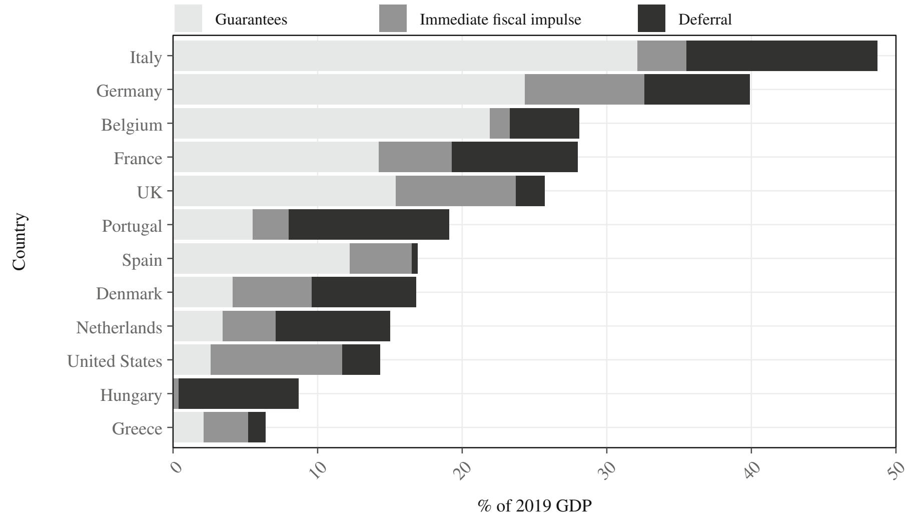


Fig. 6 COVID-19 liquidity support through fiscal policy measures by international comparison. Note: Calculations are retrieved from Anderson and et al. (2020). Numbers reflect the amount (as share of 2019 GDP) of fiscal policy measures to address adverse COVID-19 impacts on companies for selected OECD countries. Numbers are as of November 18, 2020

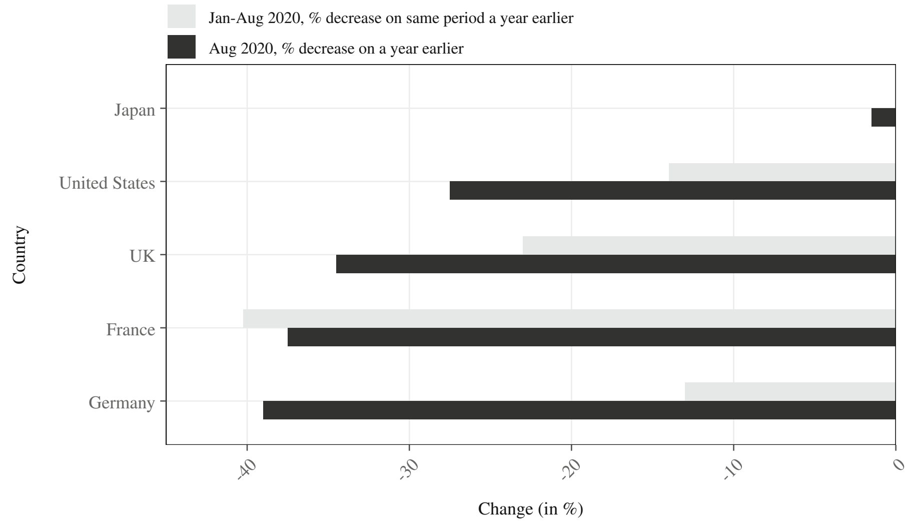


Fig. 7 Decline in corporate insolvencies during the COVID-19 crisis. Note: The figure shows the percentage change of insolvencies in the crisis year 2020 compared to 2019 for a number of selected countries. Bar chart is adapted from The Economist (2020b)

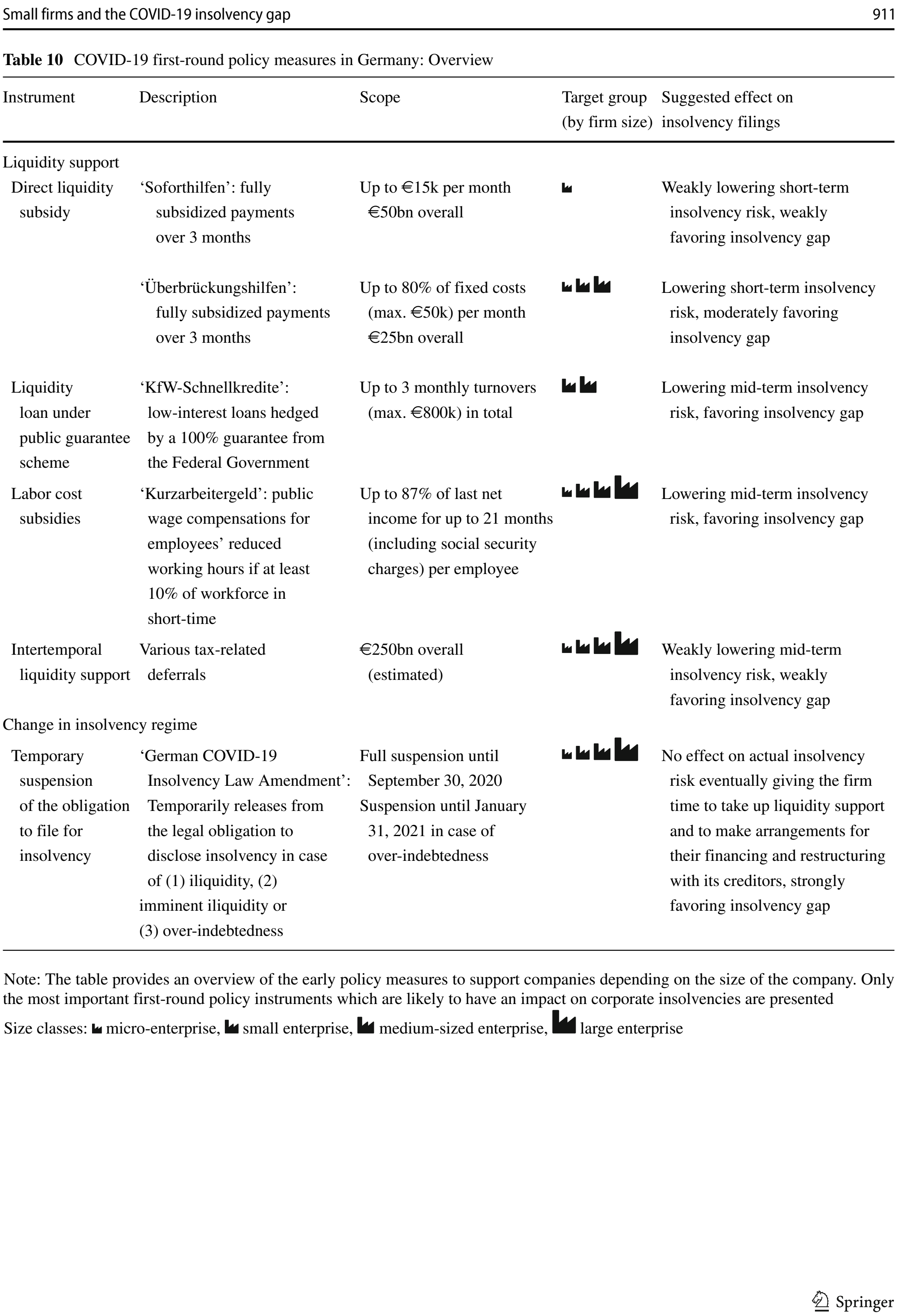


Table 10 COVID-19 first-round policy measures in Germany: Overview

Instrument	Description	Scope	Target group (by firm size)	Suggested effect on insolvency filings
Liquidity support				
Direct liquidity subsidy	‘Soforthilfen’: fully subsidized payments over 3 months	Up to €15k per month €50bn overall	🏭	Weakly lowering short-term insolvency risk, weakly favoring insolvency gap
	‘Überbrückungshilfen’: fully subsidized payments over 3 months	Up to 80% of fixed costs (max. €50k) per month €25bn overall	🏭🏭🏭	Lowering short-term insolvency risk, moderately favoring insolvency gap
Liquidity loan under public guarantee scheme	‘KfW-Schnellkredite’: low-interest loans hedged by a 100% guarantee from the Federal Government	Up to 3 monthly turnovers (max. €800k) in total	🏭🏭	Lowering mid-term insolvency risk, favoring insolvency gap
Labor cost subsidies	‘Kurzarbeitergeld’: public wage compensations for employees’ reduced working hours if at least 10% of workforce in short-time	Up to 87% of last net income for up to 21 months (including social security charges) per employee	🏭🏭🏭🏭	Lowering mid-term insolvency risk, favoring insolvency gap
Intertemporal liquidity support	Various tax-related deferrals	€250bn overall (estimated)	🏭🏭🏭🏭	Weakly lowering mid-term insolvency risk, weakly favoring insolvency gap
Change in insolvency regime				
Temporary suspension of the obligation to file for insolvency	‘German COVID-19 Insolvency Law Amendment’: Temporarily releases from the legal obligation to disclose insolvency in case of (1) iliquidity, (2) imminent iliquidity or (3) over-indebtedness	Full suspension until September 30, 2020 Suspension until January 31, 2021 in case of over-indebtedness	🏭🏭🏭🏭	No effect on actual insolvency risk eventually giving the firm time to take up liquidity support and to make arrangements for their financing and restructuring with its creditors, strongly favoring insolvency gap

Note: The table provides an overview of the early policy measures to support companies depending on the size of the company. Only the most important first-round policy instruments which are likely to have an impact on corporate insolvencies are presented

Size classes: 🏭 micro-enterprise, 🏭 small enterprise, 🏭 medium-sized enterprise, 🏭 large enterprise

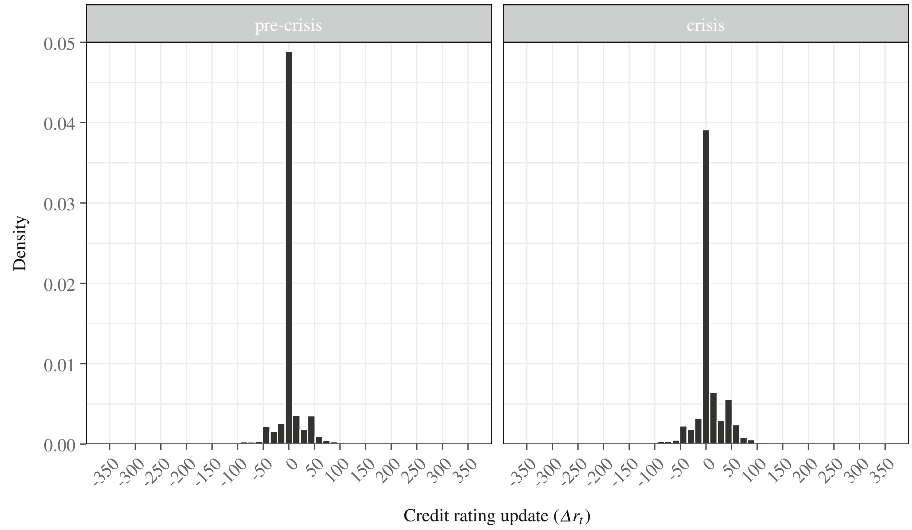


Fig. 8 Distribution of credit rating update in pre-crisis and crisis period. We define a rating update as a reassessment of the company's creditworthiness performed by Creditreform. We have precise information on the date of reassessment, which allows us to accurately assign the update to either the pre-crisis or the crisis period and also to accurately match the updates with insolvency dates. It should also be noted that a reassessment does not necessarily lead to a change in the rating index. If the creditworthiness of the company has not changed since the last rating, the company gets assigned the same index as before, resulting in a value of 0 in Δr_t. Form the figure it becomes apparent that there is a rightward shift in the distribution of rating updates during the crisis period, indicating that there were more credit rating downgrades as compared to the pre-pandemic period. This reflects that the financial situation deteriorated for a larger share of companies in the crisis period than in the three years preceding the crisis

Table 11 Calculation of the insolvency gap in absolute terms

Sector	Size of company						$\sum$
	Micro		Small		Medium		
	N_s	IG_s (in %)	N_s	IG_s (in %)	N_s	IG_s (in %)	
Accommodation & catering	37,633	0.0115	4,852	0.0005	810	0.0028	
Creative industry & entertainment	16,057	0.0012	1,910	0.0017	476	0.0000	
Food production	8,191	0.0027	3,674	0.0024	1,962	−0.0019	
Health & social services	69,029	0.0037	12,331	0.0005	4,269	−0.0011	
Insurance & banking	46,670	0.0037	2,583	0.0000	1,290	0.0000	
Logistics & transport	43,899	0.0070	10,756	0.0002	2,773	0.0030	
Chemicals & pharmaceuticals	5,170	0.0033	3,980	0.0003	2,342	0.0000	

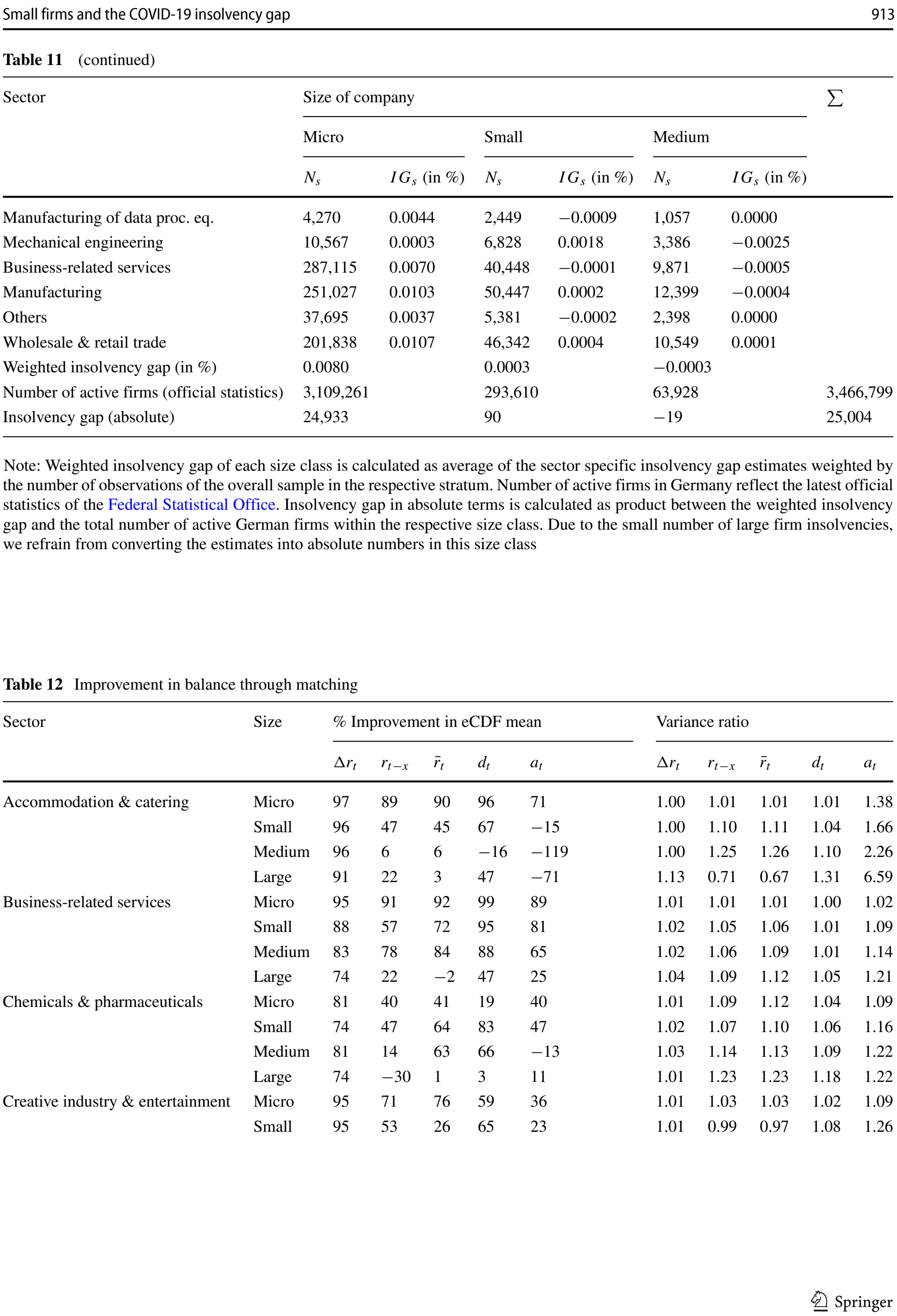


Table 11 (continued)

Sector	Size of company						$\sum$
	Micro		Small		Medium		
	N_s	IG_s (in %)	N_s	IG_s (in %)	N_s	IG_s (in %)	
Manufacturing of data proc. eq.	4,270	0.0044	2,449	−0.0009	1,057	0.0000	
Mechanical engineering	10,567	0.0003	6,828	0.0018	3,386	−0.0025	
Business-related services	287,115	0.0070	40,448	−0.0001	9,871	−0.0005	
Manufacturing	251,027	0.0103	50,447	0.0002	12,399	−0.0004	
Others	37,695	0.0037	5,381	−0.0002	2,398	0.0000	
Wholesale & retail trade	201,838	0.0107	46,342	0.0004	10,549	0.0001	
Weighted insolvency gap (in %)	0.0080		0.0003		−0.0003		
Number of active firms (official statistics)	3,109,261		293,610		63,928		3,466,799
Insolvency gap (absolute)	24,933		90		−19		25,004

Note: Weighted insolvency gap of each size class is calculated as average of the sector specific insolvency gap estimates weighted by the number of observations of the overall sample in the respective stratum. Number of active firms in Germany reflect the latest official statistics of the Federal Statistical Office. Insolvency gap in absolute terms is calculated as product between the weighted insolvency gap and the total number of active German firms within the respective size class. Due to the small number of large firm insolvencies, we refrain from converting the estimates into absolute numbers in this size class

Table 12 Improvement in balance through matching

Sector	Size	% Improvement in eCDF mean					Variance ratio				
		Δr_t	r_{t-x}	$\bar{r}_t$	d_t	a_t	Δr_t	r_{t-x}	$\bar{r}_t$	d_t	a_t
Accommodation & catering	Micro	97	89	90	96	71	1.00	1.01	1.01	1.01	1.38
	Small	96	47	45	67	−15	1.00	1.10	1.11	1.04	1.66
	Medium	96	6	6	−16	−119	1.00	1.25	1.26	1.10	2.26
	Large	91	22	3	47	−71	1.13	0.71	0.67	1.31	6.59
Business-related services	Micro	95	91	92	99	89	1.01	1.01	1.01	1.00	1.02
	Small	88	57	72	95	81	1.02	1.05	1.06	1.01	1.09
	Medium	83	78	84	88	65	1.02	1.06	1.09	1.01	1.14
	Large	74	22	−2	47	25	1.04	1.09	1.12	1.05	1.21
Chemicals & pharmaceuticals	Micro	81	40	41	19	40	1.01	1.09	1.12	1.04	1.09
	Small	74	47	64	83	47	1.02	1.07	1.10	1.06	1.16
	Medium	81	14	63	66	−13	1.03	1.14	1.13	1.09	1.22
	Large	74	−30	1	3	11	1.01	1.23	1.23	1.18	1.22
Creative industry & entertainment	Micro	95	71	76	59	36	1.01	1.03	1.03	1.02	1.09
	Small	95	53	26	65	23	1.01	0.99	0.97	1.08	1.26

Table 12 (continued)

Sector	Size	% Improvement in eCDF mean					Variance ratio				
		Δr_t	r_{t-x}	$\bar{r}_t$	d_t	a_t	Δr_t	r_{t-x}	$\bar{r}_t$	d_t	a_t
	Medium	94	−71	−64	−3	−79	1.03	1.29	1.33	1.11	2.15
	Large	87	1	19	87	−19	1.07	0.81	0.72	1.05	0.60
Food production	Micro	85	77	79	67	32	1.01	1.04	1.03	1.04	1.16
	Small	83	36	42	89	−28	1.00	1.03	1.03	1.02	1.18
	Medium	71	−50	−19	12	−58	1.03	1.28	1.29	1.07	1.14
	Large	62	55	47	33	34	1.00	1.48	1.48	1.22	1.02
Health & social services	Micro	95	89	91	92	80	1.01	1.01	1.01	1.01	1.18
	Small	89	47	50	25	35	1.02	1.05	1.05	1.03	1.17
	Medium	84	42	24	78	30	1.01	1.12	1.12	1.08	1.11
	Large	71	54	39	79	−39	1.03	1.24	1.27	1.10	1.44
Insurance & banking	Micro	90	86	88	89	80	1.01	1.02	1.02	1.00	1.05
	Small	73	49	67	82	73	1.03	1.04	1.04	1.07	1.07
	Medium	52	61	50	92	63	1.04	1.05	1.07	1.00	1.06
	Large	79	59	59	96	71	1.08	1.17	1.16	1.02	0.98
Logistics & transport	Micro	93	87	87	93	62	1.01	1.02	1.02	1.01	1.05
	Small	89	−27	49	0	53	1.02	1.12	1.13	1.04	1.09
	Medium	84	−9	35	52	11	1.02	1.14	1.16	1.07	1.19
	Large	85	−78	−107	77	−9	1.05	1.30	1.26	1.09	1.26
Manufacturing	Micro	93	92	93	97	82	1.01	1.01	1.01	1.00	1.03
	Small	87	72	80	99	68	1.02	1.04	1.05	1.00	1.04
	Medium	84	−2	54	67	37	1.02	1.09	1.11	1.03	1.07
	Large	85	−27	19	73	−23	1.03	1.23	1.20	1.12	1.25
Manufacturing of data processing equipment	Micro	74	−20	0	89	44	1.01	1.06	1.05	1.05	1.19
	Small	71	19	43	81	32	1.01	1.20	1.16	1.07	1.28
	Medium	73	−31	23	−35	45	1.05	1.55	1.58	1.11	1.27
	Large	79	32	56	65	3	1.04	1.11	1.04	1.17	1.60
Mechanical engineering	Micro	83	27	31	92	45	1.01	1.07	1.06	1.01	1.11
	Small	77	−9	37	89	25	1.02	1.10	1.13	1.02	1.12
	Medium	85	−22	37	78	−7	1.01	1.22	1.18	1.05	1.19
	Large	87	3	18	54	28	1.02	1.30	1.39	1.08	1.16
Others	Micro	93	88	90	91	64	1.01	1.01	1.02	1.01	1.12
	Small	86	32	52	93	35	1.02	1.02	1.04	1.01	1.26
	Medium	84	45	63	72	42	1.03	1.08	1.08	1.02	1.05
	Large	74	25	48	80	−47	1.01	1.08	1.06	1.12	1.11
Wholesale & retail trade	Micro	96	91	92	98	82	1.01	1.01	1.01	1.00	1.03
	Small	92	11	49	93	67	1.01	1.04	1.04	1.01	1.04
	Medium	87	33	66	82	42	1.02	1.11	1.11	1.02	1.10
	Large	83	30	45	75	20	1.03	1.18	1.17	1.08	1.15

Note: The table shows balance assessment statistics for all matching variables. % improvement in empirical cumulative density function (eCDF) mean shows by how much percent the deviation in the eCDF mean between pre-crisis and crisis observations has improved through nearest neighbor matching. It becomes apparent that for most covariates in all sector-size strata a substantial improvement in balance has been achieved through the matching process. Variance ratio statistics refer to the ratio of the variance among the matched control observations and the variance among the crisis observations for the respective variable. Values closer to zero indicate better balance in variance

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Acs, Z. J., Desai, S., & Hessels, J. (2008). Entrepreneurship, economic development and institutions. *Small Business Economics*, 31(3), 219–234. https://doi.org/10.1007/s11187-008-9135-9.

Adalet McGowan, M., Andrews, D., & Millot, V. (2018). The walking dead? Zombie firms and productivity performance in OECD countries. *Economic Policy*, 33(96), 685–736. https://doi.org/10.1093/epolic/eiy012.

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8), 1541–1551. https://doi.org/10.1016/j.jbankfin.2007.07.014.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the Z-Score and ZETA® models. In A.R. Bell, C. Brooks, & M. Prokopczuk (Eds.) *Handbook of research methods and applications in empirical finance* (pp. 428–456). Edward Elgar Publishing. https://doi.org/10.4337/9780857936097.00027.

Anderson, J., et al. (2020). The fiscal response to the economic fallout from the coronavirus. Accessed: 28 Dec 2020. https://www.bruegel.org/publications/datasets/covid-national-dataset/.

Archibugi, D., Filippetti, A., & Frenz, M. (2013). Economic crisis and innovation: is destruction prevailing over accumulation? *Research Policy*, 42(2), 303–314. https://doi.org/10.1016/j.respol.2012.07.002.

Arcuri, G., & Levratto, N. (2020). Early stage SME bankruptcy: does the local banking market matter? *Small Business Economics*, 54(2), 421–436. https://doi.org/10.1007/s11187-018-0042-4.

Barrero, J. M., Bloom, N., & Davis, S. (2020). COVID-19 Is Also a Reallocation Shock. NBER Working Paper 27137. https://doi.org/10.3386/w27137.

Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., & Stanton, C. (2020). The impact of COVID-19 on small business outcomes and expectations. *Proceedings of the National Academy of Sciences*, 117(30), 17656–17666. https://doi.org/10.1073/pnas.2006991117.

Baumol, W. J. (1990). Entrepreneurship: productive, Unproductive, and Destructive. *Journal of Political Economy*, 98(5, Part 1), 893–921. https://doi.org/10.1086/261712.

Berger, A. N., Cerqueiro, G., & Penas, M. F. (2011). Does debtor protection really protect debtors? Evidence from the small business credit market. *Journal of Banking and Finance*, 35(7), 1843–1857. https://doi.org/10.1016/12.010j.jbankfin.2010.

Bersch, J., Gottschalk, S., Mueller, B., & Niefert, M. (2014). The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. ZEW Discussion Paper, 14–104. https://doi.org/10.2139/ssrn.2548385.

Block, J. H., Fisch, C., & Hirschmann, M. (2021). The determinants of bootstrap financing in crises: evidence from entrepreneurial ventures in the COVID-19 pandemic. *Small Business Economics*, (forthcoming). https://doi.org/10.1007/s11187-020-00445-6.

Brouwer, M. (2006). Reorganization in US and European bankruptcy law. *European Journal of Law and Economics*, 22(1), 5–20. https://doi.org/10.1007/s10657-006-8978-2.

Caballero, R. J., & Hammour, M. L. (1994). The cleansing effect of recessions. *American Economic Review*, 84(5), 1350–1368. http://www.jstor.org/stable/2117776.

Caballero, R. J., Hoshi, T., & Kashyap, A. K. (2008). Zombie lending and depressed restructuring in Japan. *American Economic Review*, 98(5), 1943–1977. https://doi.org/10.1257/aer.98.5.1943.

Cahuc, P., Kramarz, F., & Nevoux, S. (2018). When short-time work works. *Banque de France Working Paper*, 692. https://doi.org/10.2139/ssrn.3247486.

Carreira, C., & Teixeira, P. (2016). Entry and exit in severe recessions: lessons from the 2008–2013 Portuguese economic crisis. *Small Business Economics*, 46(4), 591–617. https://doi.org/10.1007/s11187-016-9703-3.

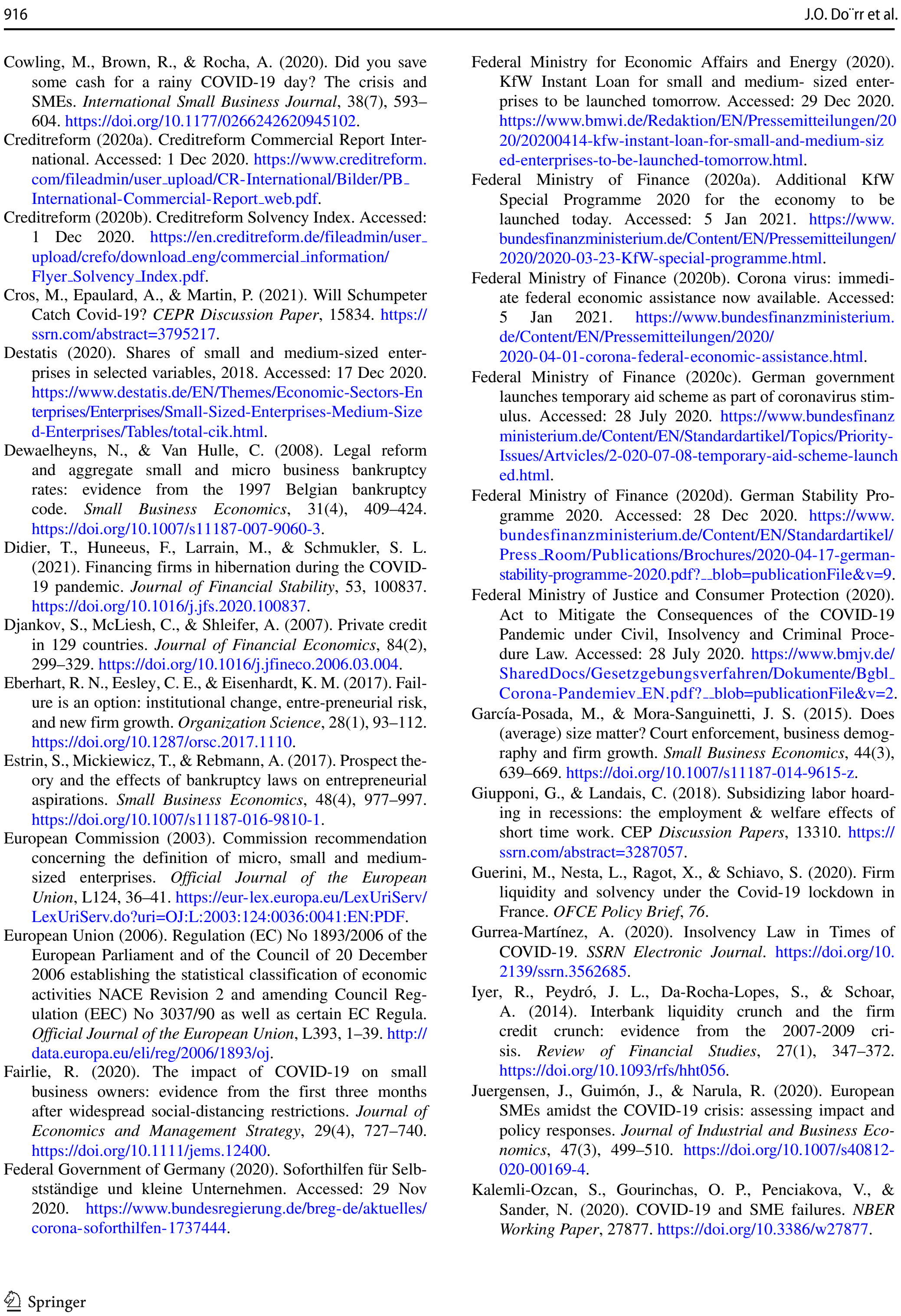
Chemin, M. (2009). The impact of the judiciary on entrepreneurship: evaluation of Pakistan's "Access to Justice Programme". *Journal of Public Economics*, 93(1–2), 114–125. https://doi.org/10.1016/j.jpubeco.2008.05.005.

Chowdhury, F., Terjesen, S., & Audretsch, D. (2015). Varieties of entrepreneurship: institutional drivers across entrepreneurial activity and country. *European Journal of Law and Economics*, 40(1), 121–148. https://doi.org/10.1007/s10657-014-9464-x.

Cook, G. A., Pandit, N. R., & Milman, D. (2001). Formal rehabilitation procedures and insolvent firms: empirical evidence on the British Company voluntary arrangement procedure. *Small Business Economics*, 17(4), 255–271. https://doi.org/10.1023/A:1012293605945.

Council of the European Union (2020). Report on the comprehensive economic policy response to the COVID-19 pandemic. Accessed: 12 Dec 2020. https://www.consilium.europa.eu/en/press/press-releases/2020/04/09/report-on-the-comprehensive-economic-policy-response-to-the-covid-19-pandemic/.

Cowling, M., Liu, W., & Ledger, A. (2012). Small business financing in the UK before and during the current financial crisis. *International Small Business Journal*, 30(7), 778–800. https://doi.org/10.1177/0266242611435516.



Cowling, M., Brown, R., & Rocha, A. (2020). Did you save some cash for a rainy COVID-19 day? The crisis and SMEs. *International Small Business Journal*, 38(7), 593–604. https://doi.org/10.1177/0266242620945102.

Creditreform (2020a). Creditreform Commercial Report International. Accessed: 1 Dec 2020. https://www.creditreform.com/fileadmin/user_upload/CR-International/Bilder/PB_International-Commercial-Report_web.pdf.

Creditreform (2020b). Creditreform Solvency Index. Accessed: 1 Dec 2020. https://en.creditreform.de/fileadmin/user_upload/crefo/download_eng/commercial_information/Flyer_Solvency_Index.pdf.

Cros, M., Epaulard, A., & Martin, P. (2021). Will Schumpeter Catch Covid-19? *CEPR Discussion Paper*, 15834. https://ssrn.com/abstract=3795217.

Destatis (2020). Shares of small and medium-sized enterprises in selected variables, 2018. Accessed: 17 Dec 2020. https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Enterprises/Small-Sized-Enterprises-Medium-Sized-Enterprises/Tables/total-cik.html.

Dewaelheyns, N., & Van Hulle, C. (2008). Legal reform and aggregate small and micro business bankruptcy rates: evidence from the 1997 Belgian bankruptcy code. *Small Business Economics*, 31(4), 409–424. https://doi.org/10.1007/s11187-007-9060-3.

Didier, T., Huneeus, F., Larrain, M., & Schmukler, S. L. (2021). Financing firms in hibernation during the COVID-19 pandemic. *Journal of Financial Stability*, 53, 100837. https://doi.org/10.1016/j.jfs.2020.100837.

Djankov, S., McLiesh, C., & Shleifer, A. (2007). Private credit in 129 countries. *Journal of Financial Economics*, 84(2), 299–329. https://doi.org/10.1016/j.jfineco.2006.03.004.

Eberhart, R. N., Eesley, C. E., & Eisenhardt, K. M. (2017). Failure is an option: institutional change, entre-preneurial risk, and new firm growth. *Organization Science*, 28(1), 93–112. https://doi.org/10.1287/orsc.2017.1110.

Estrin, S., Mickiewicz, T., & Rebmann, A. (2017). Prospect theory and the effects of bankruptcy laws on entrepreneurial aspirations. *Small Business Economics*, 48(4), 977–997. https://doi.org/10.1007/s11187-016-9810-1.

European Commission (2003). Commission recommendation concerning the definition of micro, small and medium-sized enterprises. *Official Journal of the European Union*, L124, 36–41. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:124:0036:0041:EN:PDF.

European Union (2006). Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regula. *Official Journal of the European Union*, L393, 1–39. http://data.europa.eu/eli/reg/2006/1893/oj.

Fairlie, R. (2020). The impact of COVID-19 on small business owners: evidence from the first three months after widespread social-distancing restrictions. *Journal of Economics and Management Strategy*, 29(4), 727–740. https://doi.org/10.1111/jems.12400.

Federal Government of Germany (2020). Soforthilfen für Selbstständige und kleine Unternehmen. Accessed: 29 Nov 2020. https://www.bundesregierung.de/breg-de/aktuelles/corona-soforthilfen-1737444.

Federal Ministry for Economic Affairs and Energy (2020). KfW Instant Loan for small and medium- sized enterprises to be launched tomorrow. Accessed: 29 Dec 2020. https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2020/20200414-kfw-instant-loan-for-small-and-medium-sized-enterprises-to-be-launched-tomorrow.html.

Federal Ministry of Finance (2020a). Additional KfW Special Programme 2020 for the economy to be launched today. Accessed: 5 Jan 2021. https://www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2020/2020-03-23-KfW-special-programme.html.

Federal Ministry of Finance (2020b). Corona virus: immediate federal economic assistance now available. Accessed: 5 Jan 2021. https://www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2020/2020-04-01-corona-federal-economic-assistance.html.

Federal Ministry of Finance (2020c). German government launches temporary aid scheme as part of coronavirus stimulus. Accessed: 28 July 2020. https://www.bundesfinanzministerium.de/Content/EN/Standardartikel/Topics/Priority-Issues/Artvicles/2-020-07-08-temporary-aid-scheme-launched.html.

Federal Ministry of Finance (2020d). German Stability Programme 2020. Accessed: 28 Dec 2020. https://www.bundesfinanzministerium.de/Content/EN/Standardartikel/Press_Room/Publications/Brochures/2020-04-17-german-stability-programme-2020.pdf?__blob=publicationFile&v=9.

Federal Ministry of Justice and Consumer Protection (2020). Act to Mitigate the Consequences of the COVID-19 Pandemic under Civil, Insolvency and Criminal Procedure Law. Accessed: 28 July 2020. https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/Bgbl_Corona-Pandemiev_EN.pdf?__blob=publicationFile&v=2.

García-Posada, M., & Mora-Sanguinetti, J. S. (2015). Does (average) size matter? Court enforcement, business demography and firm growth. *Small Business Economics*, 44(3), 639–669. https://doi.org/10.1007/s11187-014-9615-z.

Giupponi, G., & Landais, C. (2018). Subsidizing labor hoarding in recessions: the employment & welfare effects of short time work. CEP *Discussion Papers*, 13310. https://ssrn.com/abstract=3287057.

Guerini, M., Nesta, L., Ragot, X., & Schiavo, S. (2020). Firm liquidity and solvency under the Covid-19 lockdown in France. *OFCE Policy Brief*, *76*.

Gurrea-Martínez, A. (2020). Insolvency Law in Times of COVID-19. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3562685.

Iyer, R., Peydró, J. L., Da-Rocha-Lopes, S., & Schoar, A. (2014). Interbank liquidity crunch and the firm credit crunch: evidence from the 2007-2009 crisis. *Review of Financial Studies*, 27(1), 347–372. https://doi.org/10.1093/rfs/hht056.

Juergensen, J., Guimón, J., & Narula, R. (2020). European SMEs amidst the COVID-19 crisis: assessing impact and policy responses. *Journal of Industrial and Business Economics*, 47(3), 499–510. https://doi.org/10.1007/s40812-020-00169-4.

Kalemli-Ozcan, S., Gourinchas, O. P., Penciakova, V., & Sander, N. (2020). COVID-19 and SME failures. *NBER Working Paper*, 27877. https://doi.org/10.3386/w27877.

Kopp, D., & Siegenthaler, M. (2021). Short-time work in and after the Great Recession. *Journal of the European Economic Association* (forthcoming). https://doi.org/10.1093/jeea/jvab003.

Lee, N., Sameen, H., & Cowling, M. (2015). Access to finance for innovative SMEs since the financial crisis. *Research Policy*, 44(2), 370–380. https://doi.org/10.1016/j.respol.2014.09.008.

Lee, S. H., Yamakawa, Y., Peng, M. W., & Barney, J. B. (2011). How do bankruptcy laws affect entrepreneurship development around the world? *Journal of Business Venturing*, 26(5), 505–520. https://doi.org/10.1016/j.jbusvent.2010.05.001.

Legrand, M. D. P. (2017). Retrospectives: do productive recessions show the recuperative powers of capitalism? Schumpeter's analysis of the cleansing effect. *Journal of Economic Perspectives*, 31(1), 245–256. https://doi.org/10.1257/jep.31.1.245.

Levie, J., Autio, E., Acs, Z., & Hart, M. (2014). Global entrepreneurship and institutions: an introduction. *Small Business Economics*, 42(3), 437–444. https://doi.org/10.1007/s11187-013-9516-6.

McGuinness, G., & Hogan, T. (2016). Bank credit and trade credit: evidence from SMEs over the financial crisis. *International Small Business Journal*, 34(4), 412–445. https://doi.org/10.1177/0266242614558314.

Melcarne, A., & Ramello, G. B. (2020). Bankruptcy delay and firms' dynamics. *Small Business Economics*, 54(2), 405–419. https://doi.org/10.1007/s11187-018-0041-5.

OECD (2020a). Coronavirus (COVID-19): SME policy responses. Accessed: 17 Dec 2020. https://read.oecd-ilibrary.org/view/?ref=119_119680-di6h3qgi4x&title=Covid-19_SME_Policy_Responses.

OECD (2020b). Corporate sector vulnerabilities during the Covid-19 outbreak: assessment and policy responses. Accessed: 17 Dec 2020. https://doi.org/10.1787/6434b1e4-en.

Peng, M. W., Yamakawa, Y., & Lee, S. H. (2010). Bankruptcy laws and entrepreneur- friendliness. *Entrepreneurship: Theory and Practice*, 34(3), 517–530. https://doi.org/10.1111/j.1540-6520.2009.00350.x.

Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76(374), 221–230. https://doi.org/10.1080/01621459.1981.10477633.

Rodano, G., Serrano-Velarde, N., & Tarantino, E. (2016). Bankruptcy law and bank financing. *Journal of Financial Economics*, 120(2), 363–382. https://doi.org/10.1016/j.jfineco.2016.01.016.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 29(1), 159–183. https://doi.org/10.2307/2529684.

Rubin, D. B. (1980). Bias reduction using Mahalanobis-Metric matching. *Biometrics*, 36(2), 293–298. https://doi.org/10.2307/2529981.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. New York: Harper.

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1), 1–21. https://doi.org/10.1214/09-STS313.

The Economist (2020a). The corporate undead: what to do about zombie firms. Accessed: 5 Jan 2021. https://www.economist.com/leaders/2020/09/24/what-to-do-about-zombie-firms.

The Economist (2020b). The corporate undead: why covid-19 will make killing zombie firms of harder. Accessed: 5 Jan 2021. https://www.economist.com/finance-and-economics/2020/09/26/why-covid-19-will-make-killing-zombie-firms-off-harder.

The Washington Post (2020). Here's one more economic problem the government's response to the virus has unleashed: Zombie firms. Accessed: 11 Nov 2020. https://www.washingtonpost.com/business/2020/06/23/economy-debt-coronavirus-zombie-firms/.

Wang, J., Yang, J., Iverson, B. C., & Kluender, R. (2020). Bankruptcy and the COVID-19 Crisis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3690398.

Yang, C. H., & Chen, K. H. (2009). Are small firms less efficient? *Small Business Economics*, 32(4), 375–395. https://doi.org/10.1007/s11187-007-9082-x.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix B

An Integrated Data Framework for Policy Guidance during the Coronavirus Pandemic: Towards Real-Time Decision Support for Economic Policymakers

PLOS ONE

RESEARCH ARTICLE

An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers

Julian Oliver Dörr[1,2☯]*, Jan Kinne[1,3☯], David Lenz[2,3☯], Georg Licht[1‡], Peter Winker[2‡]

1 Department of Economics of Innovation and Industrial Dynamics, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany, **2** Department of Econometrics and Statistics, Justus Liebig University Giessen, Gießen, Germany, **3** istari.ai, Mannheim, Germany

☯ These authors contributed equally to this work.
‡These authors also contributed equally to this work.
* julian.doerr@zew.de

OPEN ACCESS

Citation: Dörr JO, Kinne J, Lenz D, Licht G, Winker P (2022) An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. PLoS ONE 17(2): e0263898. https://doi.org/10.1371/journal.pone.0263898

Editor: Liviu-Adrian Cotfas, Bucharest University of Economic Studies, ROMANIA

Received: July 20, 2021

Accepted: January 29, 2022

Published: February 14, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0263898

Copyright: © 2022 Dörr et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: This work uses firm-level survey and credit rating data which cannot be published. The data is available at the ZEW Data Research Centre, http://kooperationen.zew.de/en/

Abstract

Usually, official and survey-based statistics guide policymakers in their choice of response instruments to economic crises. However, in an early phase, after a sudden and unforeseen shock has caused unexpected and fast-changing dynamics, data from traditional statistics are only available with non-negligible time delays. This leaves policymakers uncertain about how to most effectively manage their economic countermeasures to support businesses, especially when they need to respond quickly, as in the COVID-19 pandemic. Given this information deficit, we propose a framework that guided policymakers throughout all stages of this unforeseen economic shock by providing timely and reliable sources of firm-level data as a basis to make informed policy decisions. We do so by combining early stage 'ad hoc' web analyses, 'follow-up' business surveys, and 'retrospective' analyses of firm outcomes. A particular focus of our framework is on assessing the early effects of the pandemic, using highly dynamic and large-scale data from corporate websites. Most notably, we show that textual references to the coronavirus pandemic published on a large sample of company websites and state-of-the-art text analysis methods allowed to capture the heterogeneity of the pandemic's effects at a very early stage and entailed a leading indication on later movements in firm credit ratings. While the proposed framework is specific to the COVID-19 pandemic, the integration of results obtained from real-time online sources in the design of subsequent surveys and their value in forecasting firm-level outcomes typically targeted by policy measures, is a first step towards a more timely and holistic approach for policy guidance in times of economic shocks.

1 Introduction

COVID-19 and its economic consequences have placed numerous firms under severe distress. In almost all countries, stores and businesses were closed and mobility severely restricted to contain the spread of the virus. While these large-scale anti-contagion policies had provably

zew-fdz upon request. The website data is archived in ZEW's Dark Archive for the purpose of adherence to scientific standards and proving correct scientific work according to § 60d III p. 2 of Copy Right Act, https://www.zew.de/en/das-zew/serviceeinheiten/zentraledienstleistungen/bibliothek/. A current version of the website data may be collected using the ARGUS Webscraper (https://github.com/datawizard1337/ARGUS). The code for analyzing the data as presented in the study is available at https://github.com/julienOlivier3/DataFramework_EconomicCrises.

Funding: The study extends upon a project analyzing the economic effects on SMEs in the COVID-19 crisis which was funded by the German Federal Ministry of Economic Affairs and Energy (https://www.bmwi.de) under the grant agreement No 15/20 (GL). Funding covered the collection of webdata and the analysis of both webdata and survey data. Moreover, the project received support by the Ministry of Science, Research and the Arts of the government of Baden Wuerttemberg (https://mwk.baden-wuerttemberg.de) as part of its Science Data Center program under the grant "Business and Economic Research Data Center (BERD)" (GL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

positive effects on health outcomes [1], they fundamentally changed the landscape for many businesses. Due to the forced halt of many economic activities and the severe shock to global trade, many companies faced a situation of reduced business activity and declining sales figures, as well as major disturbances to their value chains and supplier networks, which had immediate consequences on the affected firms' financial positions.

The impact of COVID-19 on businesses has shown, however, a great degree of heterogeneity [2–4]. In some sectors firms have been barely affected by the pandemic or have even benefited from it, while in others large numbers of companies have been pushed into financial distress. Besides sector-specific differences, the economic exposure to the pandemic has also strongly varied with companies' business models. Some operations managed to adjust swiftly to the changed conditions, others had little scope to do so [5].

Now, after more than a year of pandemic, the winners and losers of the crisis seem rather clear. While firms with highly digitized business models such as delivery companies, e-commerce as well as online video conferencing and education platforms have thrived, companies whose business models are characterized by physical human interaction such as culture, travel, hospitality, restaurants and retail trade have greatly suffered [3]. What seems clear today, has however not been obvious at an early stage of the shock, when policymakers were confronted with various forms of economic uncertainty [6, 7] and stepped largely in the dark about the impacts of the pandemic on different businesses and different industries. Not only the dynamics of the pandemic were hard to foresee at an early stage of the crisis, but also governments' economic response measures have been unprecedented such that referencing to previous experiences has neither been useful nor possible. A dilemma for policymakers, who were forced to act quickly to cushion the economic impact of their virus containment measures, which severely added to the plight of businesses.

In fact, the shutdown measures not only required companies to reorganize their operations by adjusting to the changed conditions, but also led to a fast erosion of equity positions among heavily exposed companies. This brought many firms on the brink of financial solvency [8] and thus called for fast government assistance [9]. Faced with the threat of a wave of corporate insolvencies and its immediate consequences, such as mass layoffs, but also with a lack of information about the heterogeneity of the economic impact of the shock, policymakers granted liquidity subsidies and other support instruments on an unprecedented scale [9]. In Germany, the focus country of our study, for example, the government even launched the 'largest assistance package in the history of the Federal Republic of Germany' [10, pg. 3] comprising public net borrowing of around €156bn [10].

The lack of early indicators, signaling which firms were at highest risk to suffer liquidity shortfalls and indicating how sectors and regions were differently exposed in the early stage after the economic shock [9], left policymakers uncertain about how to most effectively steer countermeasures. As a result, most of the early stimulus was awarded on a lump-sum basis, without taking into account that not all companies were equally affected by the pandemic [11] (e.g. in Germany, liquidity grants' 'application and payment process [needed] to be swift and free from red tape' according to the Ministry of Finance [12, para. 2]. In the context of public loan programs, 'the credit approval process [did] not involve additional credit risk assessment by the bank' and 'there [were] no requirements for collateral security' [13, para. 5]). In this sense, the coronavirus pandemic demonstrated that in highly dynamic times, policymakers face information deficits that leave them uncertain about how to most effectively manage countermeasures [14]. Especially if quick intervention is required to mitigate social costs, as in the early stage of the pandemic, policymakers had no option but to grant economic aid in a largely indiscriminate manner, often at high fiscal burden. Overcoming these information deficits is therefore crucial for policymakers to steer their response measures more effectively into directions where help is needed most urgently while not overburdening fiscal budget.

Usually, policymakers draw their information from official and survey-based statistics to decide over economic stimulus measures. However, after a sudden and unforeseen shock caused incalculable and fast-changing dynamics, firm-level data from these traditional information sources is usually not yet available due to their rather inflexible and slow update cycle. A problem that applies in particular to information about smaller, unlisted companies [15]. Besides the lack in timeliness, surveys are also costly and typically do not serve well to study heterogeneity across regions and firm-specific subgroups given their relatively small sample sizes [16]. With no time to wait for official surveys to reveal the early effects of the sudden Corona shock, many governments have thus started to experiment with alternative real-time data sources during the pandemic to better understand its economic impact [14]. This has also called economic research, as important guide to public-sector decision-making, to integrate timelier sources of data at a granular level when consulting political decision-makers [17].

Following this call, we present an approach that tracks communication patterns on corporate websites to disambiguate the heterogeneity of the pandemic's impact at both industry and firm levels. Our approach relies on textual references to the coronavirus pandemic published by companies on their corporate websites. We refer to a coronavirus reference as self-reported text fragment (sentence or paragraph) that contains specific keywords associated to the pandemic and the SARS-CoV 2 virus. Our analyses show that companies used their websites to communicate about the pandemic in different contexts. Given the different context of the text references, it is possible to construct impact indicators that reflect in which dimension the firm is affected by the pandemic. State-of-the-art methods from the field of Natural Language Processing (NLP) allow to derive these indicators. We apply our framework to a large sample containing all economically active firms in Germany that have their own web domain. The dynamic nature of website data allowed us to provide policy-relevant insights at a very early stage and at near real-time, way before alternative sources could reveal first patterns at comparable granularity. Specifically, our results reveal strong heterogeneity, disaggregated by regions and at fine granular level of industry affiliation. Moreover, we show that the communication patterns serve as leading indicators for liquidity shortfalls at the firm-level. We argue that these early findings serve as an empirical guide for policy actors to initiate more targeted policies, in a situation where other data cannot yet provide information on the underlying dynamics.

We acknowledge that surveys and official data, despite their lack in timeliness, are nonetheless important instruments for designing medium-term to long-term responses. Therefore, we propose a broader data framework for policy guidance that incorporates data from such sources as they become available over the course of an unexpected shock. In our study, we integrate results from a consecutive questionnaire-based business survey as well as proprietary credit rating data to assist policymakers with deeper insights that go beyond the website-generated indicators. The idea of integrating early results obtained from real-time online sources in designing subsequent surveys bears the potential of a more timely and holistic approach for policy guidance that is generally applicable in times of economic shocks. This not only allows policymakers to react more swiftly and targeted but also enables the design of medium to long-term stimulus packages based on a rich set of information that has been continuously updated over all stages of the shock. In that sense, our framework focuses on bridging the information gap that arises when traditional data collection can only create policy guidance with non-negligible time delays, especially in such highly dynamic situations as the coronavirus pandemic.

The remainder of this paper is structured as follows. Section 2 provides an overview of the relevant literature. Section 3 introduces the different sources of firm-level data that we use to capture the impacts of the COVID-19 shock on German businesses at different stages of the pandemic and at different levels of granularity. The section also discusses the insights that were generated from these sources. Section 4 empirically examines and highlights the value of webdata as source to generate early indicators that reflect the impact of COVID-19 on the corporate sector. Section 5 concludes.

2 Related literature

This study contributes to the fast growing literature on the economic effects of the coronavirus pandemic. Naturally, financial markets deliver very early expectation-based insights to what extent an exogenous shock such as COVID-19 affects the corporate sector. Ding et al. [2], for example, analyze the relationship between firm characteristics and financial market reactions using stock market information from January to May 2020 for a large number of internationally traded firms. They find that especially firms that were strongly exposed to international supply chains, with comparatively weak pre-crisis financial standing and with higher ownership by hedge funds underperformed in the months after the outbreak of the pandemic. Based on U.S. stock market returns, Ramelli et al. [18] also analyze stock market performance in response to the COVID-19 shock but more strongly focus on the timing of the effects. They also find that internationally oriented companies, which have been severely affected by disruptions in world trade and have been heavily dependent on the Chinese market, performed poorly, especially at the beginning of the shock in January 2020. At a later stage, stock market reactions started to increasingly penalize companies with thin financial reserves, with consumer services seen as the hardest hit sector.

Further studies based on business surveys find that firms' survival expectations show great heterogeneity across industries and strongly depend on expectations concerning the duration of the shock's repercussions [19]. Based on a business survey conducted between March 28, 2020 and April 04, 2020, Bartik et al. [19] find that estimated survival probabilities are particularly low in arts and entertainment, personal services, the restaurant industry and in tourism and lodging. Using the US Current Population Survey, Fairlie [15] find that major industries such as construction, restaurants, hotels, transportation and other personal services experienced strong declines in the amount of active business owners in April 2020 due to the COVID-19 shock.

Central to this paper is the question to which extent alternative sources of online data (here: foremost text data retrieved from corporate websites) and novel methods to turn this raw data into valuable information (here: methods from the field of NLP) may help policymakers to make informed and evidence-based decisions in otherwise uncertain environments. With increasing amounts of (often unstructured) data available, improved computational resources and substantial advances in analytical techniques, this question has gained importance in recent years. Athey [20], for instance, argues that there are clear limits as to how sources of 'big data' and supervised learning techniques are useful for policy guidance. This is because 'there are a number of gaps between making a prediction and making a [good] decision' [20, p.483]. The former is where data-driven models clearly thrive, the latter, however, is subject to more nuanced trade-offs which are often not encrypted in data but rather require human rationalization. Clearly, this is also true for the many policy decisions that needed to be made in response to the COVID-19 shock. Weighing between shutdown measures to contain the spread of the virus and the economic damage caused by these measures is clearly such a rationalization. Likewise, granting state aid in a whatever-it-takes fashion to prevent the risk of a wave of business failures, as well as possible windfall effects if aid measures go to non-viable firms or firms which would not have required state support, is another trade-off policymakers were confronted with in the early phase of the pandemic. Arguably, no data-driven model could have predicted the corporate outcomes resulting from different policy decisions, thereby implicitly relieving politicians of active decision making. This study, however, explores how a non-traditional, large-scale data source can serve as valuable guide in politicians' decision-making process. This question is of particular concern in situations where traditional, policy-guiding sources of small data are not yet available but swift policy action is required. In this vein, we see our study as an important contribution to the discussion about the value of combining and supplementing small data sources such as surveys with large scale online sources. In the social sciences, the integration of such heterogeneous sources of information is considered to be high and has recently been documented in various studies (e.g. [21–24]). In this context, our paper contributes to a specific use case where the combination of small and big data sources allows overcoming information deficits to reduce the risk of policy errors.

Moreover, in fragile situations where social stability is at stake, the pandemic demonstrated that it is paramount for policymakers to ensure accountability and maintain public trust in their decision making processes. Among policymakers, this has led to an increasing demand for evidence-based decision making in the wake of the COVID-19 crisis [25]. In this context, our framework provided political decision-makers with empirical evidence to legitimize policy decisions in a situation of otherwise limited information.

Finally, this paper contributes to the literature that exploits webdata as useful information source to tackle research and policy issues. In fact, the use of various sources of webdata to collect timely and reliable information has gained traction in recent years. For example, webdata from social media platforms is used for event detection to get an up-to-date picture of the situation regarding major social events [26, 27] or natural disasters [28, 29]. Both applications have also policy relevance in terms of public security and crisis management. In the field of economic research, data from company websites have also proven to be a valuable information resource. Companies typically use their websites to report on their products and services, to present their activities and reference customers, but also to inform their customers and partners about current events related to their business activities [30, 31]. Using this form of data comes, however, with a number of requirements and challenges in terms of data acquisition, data analysis and data validation. The extraction of relevant information from unstructured or semi-structured text data from corporate websites can be seen as particularly challenging here. At the same time, it promises a number of benefits, particularly in terms of granularity, timeliness, scope and cost of collection [32]. These benefits will also turn out to be key in this study. In addition to simple keyword-based approaches, e.g. to measure the diffusion of standards [33], approaches with more sophisticated NLP and Machine Learning (ML) methods in particular, have been successfully used, to generate web-based firm-level innovation indicators [34, 35], for instance.

In the following section, we will introduce a three-stage framework to analyze the impacts of the COVID-19 pandemic on the corporate sector in Germany. Special attention is paid to the first 'ad hoc' stage of our framework, in which we examined early phase pandemic-related dynamics on corporate websites for a large sample of German firms at near real-time (see also Kinne et al. [36]).

3 Multi-stage framework for crisis impact monitoring

The framework presented in this section is based on a multi-stage process that aims to provide an up-to-date and complete picture of the business landscape during the course of the unforeseen economic shock triggered by the coronavirus pandemic. At each stage, heterogeneous sources of information are used to shed light on the pandemic-related dynamics and impacts on the corporate sector. In a first 'ad hoc' stage, a monitoring system based on a systematic analysis of corporate websites is set up in the short run, which provides informative and up-to-date impact data at a very early phase and at near real-time right after the pandemic had hit the economy. In this first stage, we not only demonstrate how a dynamic stream of unstructured, digitized data can be used for the sake of updating political decision-makers when traditional forms of policy data is not available yet. We also show that indicators generated from this information source serve in forecasting how the pandemic has been materialized in the companies' performance. Based on the findings of the first stage, the second 'follow-up' stage focuses on surveys to highlight specific aspects of the crisis and their effects on businesses. Here, the knowledge gained in the first stage enables the design of more targeted surveys. In a final 'retrospective' stage, data that has only become available after the shock has materialized in the economy is used to determine the more structural impacts on firm outcomes. The objective of the proposed framework is to provide, first and foremost, decision support for economic policy. Thus, the third stage focuses on outcome variables that are typically targeted by policymakers. In this study, we focus on how the pandemic-related liquidity shortfalls materialized in companies' observed creditworthiness. Firms' creditworthiness is a key determinant for access to external funding, which was severely impaired for many companies as a result of the lockdown measures and therefore necessitated state-financed liquidity support.

A key focus of the proposed framework is timeliness of information to ensure that policymakers can justify their decisions on an empirical basis at every stage of an economic crisis. In a highly dynamic situation such as the coronavirus pandemic, where policymakers are forced to react swiftly, timeliness is a key criterion. That is why alternative sources of *timely* and *reliable* data for policy are particularly important in order to assist policymakers in designing ad hoc support measures. In this context, the idea to complement real-time online sources with traditional information sources provides a holistic approach to continuously update economic policymakers throughout all stages of any economic shock. Fig 1 gives a conceptual overview of the proposed framework and the data bases involved.

In the following, we will present the data, the methods as well as the impact results for all three stages of the framework. Finally, we outline how our proposed framework can be extended to be more generally applicable beyond the coronavirus case presented in this paper.

3.1 First stage: Ad hoc web-based impact analysis

Especially in the early weeks of the pandemic, the impact on and response of firms and in particular the heterogeneity across different economic subsectors and regions have been quite unclear until surveys and other official data revealed first insights. We filled this information gap by making use of 'COVID-19'-related announcements found on corporate websites. For this purpose, in the first stage of our framework, we have accessed corporate websites of about 1.18 million individual German companies from mid March 2020 to end of May 2020 twice a week and searched for references related to the pandemic. We have used micro data from the Mannheim Enterprise Panel (MUP) which contains information on all economically active German firms in late 2019 including their corporate web addresses [37]. Based on a labeled sample of these references, impact indicators that reflect in which context the companies reported about the pandemic have been modeled. This approach allowed to capture first patterns regarding the effects of the economic shock on corporations and its heterogeneity across different economic sectors. In the following, we will describe how we proceeded in capturing COVID-19 references from company websites and how we turned these text fragments into meaningful indicators.

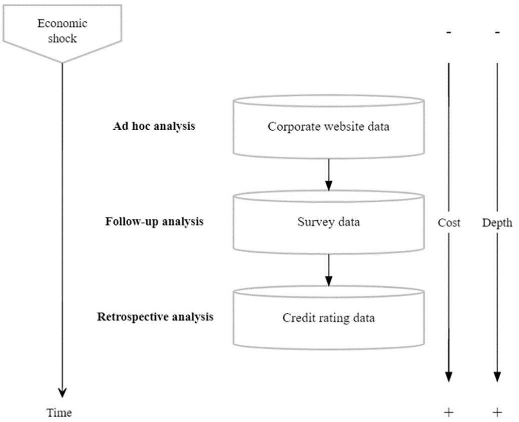


Fig 1. Framework visualization. Note: This figure illustrates the data framework for tracking effects of economic shocks on businesses by combining corporate website, business survey and credit rating data.

https://doi.org/10.1371/journal.pone.0263898.g001

In a first step, the companies' websites were queried and downloaded following a structured approach. For each corporate website address, a maximum of five webpages per company (a website usually consists of several webpages) were crawled. The selection of these webpages was not conducted at random, but followed a clear heuristic: first, webpages with the shortest Uniform Resource Locator (URL) within the corporate website domain and whose content is written in German were selected (see Kinne et al. [32] for more details on the scraping framework). The former selection criteria satisfy that those webpages with more general and up-to-date ('top-level') information were downloaded with priority making it more likely that recent Corona references were captured by the search query. The downloaded webpages were then searched for variations of the term 'COVID-19' and relevant synonyms (see S1 Table for a list of these search terms). If any of the pre-defined search terms matched, the respective Hyper-Text Markup Language (HTML) node was retrieved for further processing. This simple approach allowed for a first estimation of the number of companies reporting about the Corona pandemic on their websites as displayed in Fig 2.

In total, we queried the large sample of 1.18 million corporate websites 13 times at regular intervals during the first weeks of the COVID-19 crisis in Germany. Fig 2 reveals that at this very early stage of the outbreak, just three days after the German Federal Government announced the first nationwide economic shutdown on March 16, 2020, more than 110,000 German companies had already mentioned COVID-19 on their websites. This comprises close to 10% of the overall corporate website addresses available to us (see S2 Table for a decomposition of detected Corona references across sectors and firm sizes; S3 and S4 Tables provide detailed information on sector and firm size definitions used in this study). The growth figures in Fig 2 (red line) also show that, especially at the beginning of the pandemic, shortly after the first shutdown in Germany, information on company websites posed a highly dynamic source of crisis-related data. Within just a few days, the number of companies with Corona references grew by double digits in percentage terms. These figures suggest that corporate website content offers great potential for learning how companies are affected by the pandemic and how they are dynamically coping with the changing economic reality.

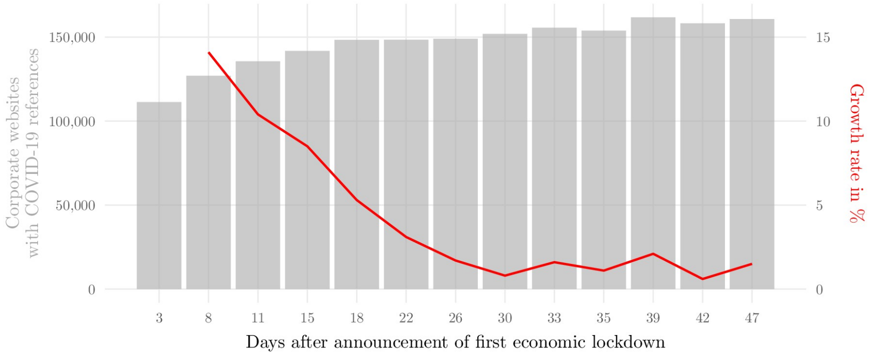


Fig 2. Companies with COVID-19 references on their corporate websites after announcement of first economic shutdown. Note: Figure shows the number of firms which reported about COVID-19 on their corporate websites over time, shortly after the announcement of the first nationwide shutdown at March 16, 2020 (left vertical axis). The repeated design of the web queries allowed to monitor the near real-time impact of the pandemic on the corporate sector. Red line (right vertical axis) depicts the growth rate of companies reporting about COVID-19 on their websites. Growth rate is calculated on a rolling basis with window size 3. Fluctuations towards the last few web queries both reflect an improved scraping process that was implemented in early May 2020 and companies that have removed COVID-19 references from their websites.

https://doi.org/10.1371/journal.pone.0263898.g002

After the relevant text passages on company websites were identified, we continued in a second step with the classification of the context of the found Corona references. To this end, we introduced five different context classes (which we were able to identify after an exploratory analysis of the references). These context categories are defined as follows:

(1) **Problem**: The company reports on problems related to the Corona pandemic. This includes but is not exclusive to closures of stores, cancellations and postponements of events, reports of delivery bottlenecks and short-time work.

(2) **No problem**: The company reports that it is not affected by the Corona pandemic or that it has no impact on its operations.

(3) **Adaption**: The company reports that it is adapting to the new circumstances. This includes measures such as new hygiene regulations, changed opening hours, home office regulations and the like.

(4) **Information**: The company reports generally, not necessarily in a business-context, about the Corona pandemic. This comprises general information about the spread of the virus, symptoms of the disease, news about the pandemic or the announcement of official regulations.

(5) **Unclear**: This group includes texts that cannot be clearly assigned. Either they are artefacts or the reference does not come with further clearly distinguishable content.

In S5 Table, the interested reader finds examples for each of the five context categories.

With already more than 250,000 distinct 'Corona' references found in the first query wave in mid March, we have made use of a pre-trained language model from the transformers family [38] to scale the context classification task. Specifically, we adapted the XLM-RoBERTa architecture [39], a multilingual transformer model [38] pre-trained on over 100 languages. XLM-RoBERTa extends upon the seminal work on Bidirectional Encoder Representations

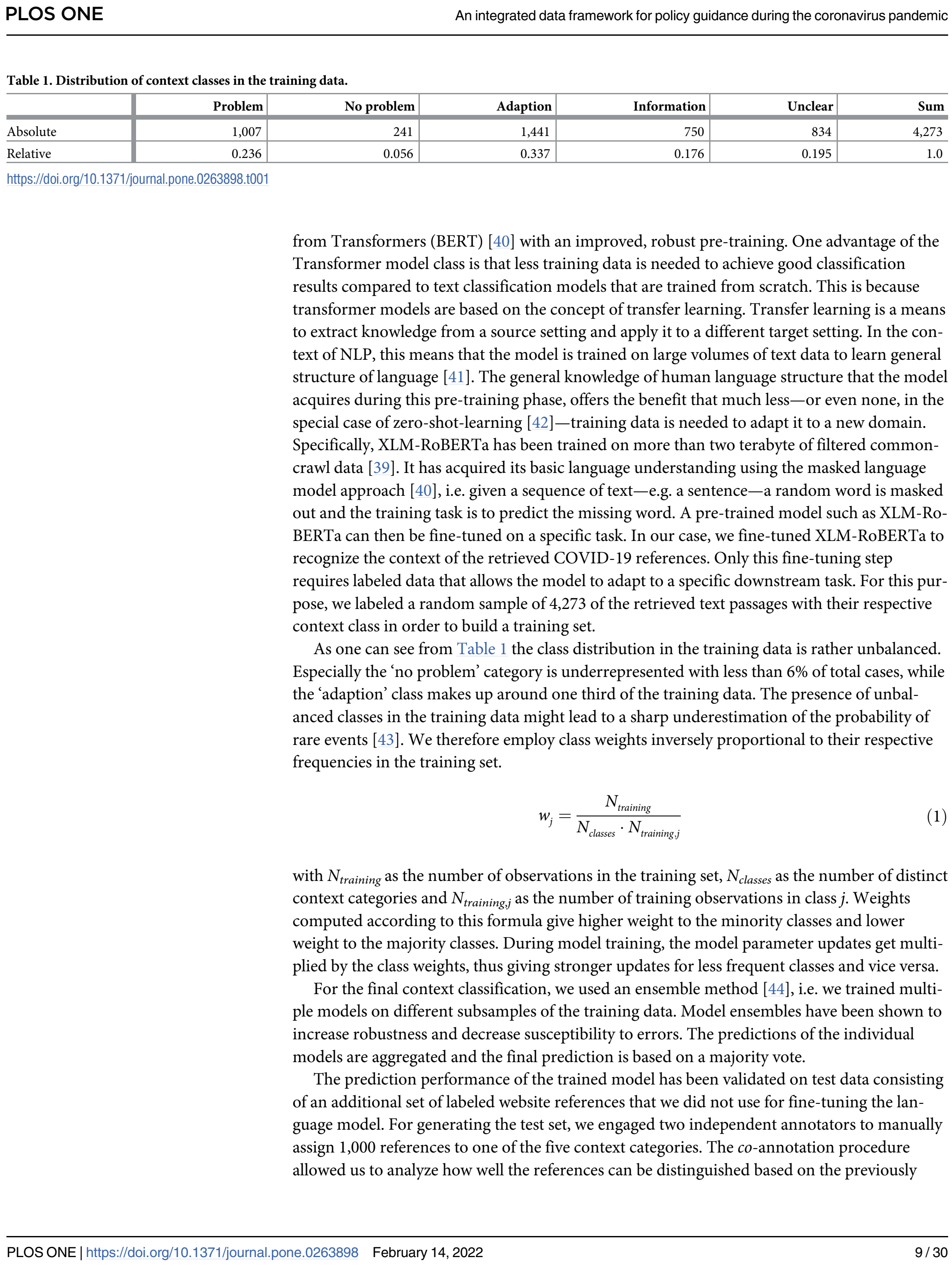


Table 1. Distribution of context classes in the training data.

	Problem	No problem	Adaption	Information	Unclear	Sum
Absolute	1,007	241	1,441	750	834	4,273
Relative	0.236	0.056	0.337	0.176	0.195	1.0

https://doi.org/10.1371/journal.pone.0263898.t001

from Transformers (BERT) [40] with an improved, robust pre-training. One advantage of the Transformer model class is that less training data is needed to achieve good classification results compared to text classification models that are trained from scratch. This is because transformer models are based on the concept of transfer learning. Transfer learning is a means to extract knowledge from a source setting and apply it to a different target setting. In the context of NLP, this means that the model is trained on large volumes of text data to learn general structure of language [41]. The general knowledge of human language structure that the model acquires during this pre-training phase, offers the benefit that much less—or even none, in the special case of zero-shot-learning [42]—training data is needed to adapt it to a new domain. Specifically, XLM-RoBERTa has been trained on more than two terabyte of filtered common-crawl data [39]. It has acquired its basic language understanding using the masked language model approach [40], i.e. given a sequence of text—e.g. a sentence—a random word is masked out and the training task is to predict the missing word. A pre-trained model such as XLM-RoBERTa can then be fine-tuned on a specific task. In our case, we fine-tuned XLM-RoBERTa to recognize the context of the retrieved COVID-19 references. Only this fine-tuning step requires labeled data that allows the model to adapt to a specific downstream task. For this purpose, we labeled a random sample of 4,273 of the retrieved text passages with their respective context class in order to build a training set.

As one can see from Table 1 the class distribution in the training data is rather unbalanced. Especially the 'no problem' category is underrepresented with less than 6% of total cases, while the 'adaption' class makes up around one third of the training data. The presence of unbalanced classes in the training data might lead to a sharp underestimation of the probability of rare events [43]. We therefore employ class weights inversely proportional to their respective frequencies in the training set.

$$w_j = \frac{N_{training}}{N_{classes} \cdot N_{training,j}} \qquad (1)$$

with $N_{training}$ as the number of observations in the training set, $N_{classes}$ as the number of distinct context categories and $N_{training,j}$ as the number of training observations in class j. Weights computed according to this formula give higher weight to the minority classes and lower weight to the majority classes. During model training, the model parameter updates get multiplied by the class weights, thus giving stronger updates for less frequent classes and vice versa.

For the final context classification, we used an ensemble method [44], i.e. we trained multiple models on different subsamples of the training data. Model ensembles have been shown to increase robustness and decrease susceptibility to errors. The predictions of the individual models are aggregated and the final prediction is based on a majority vote.

The prediction performance of the trained model has been validated on test data consisting of an additional set of labeled website references that we did not use for fine-tuning the language model. For generating the test set, we engaged two independent annotators to manually assign 1,000 references to one of the five context categories. The *co*-annotation procedure allowed us to analyze how well the references can be distinguished based on the previously

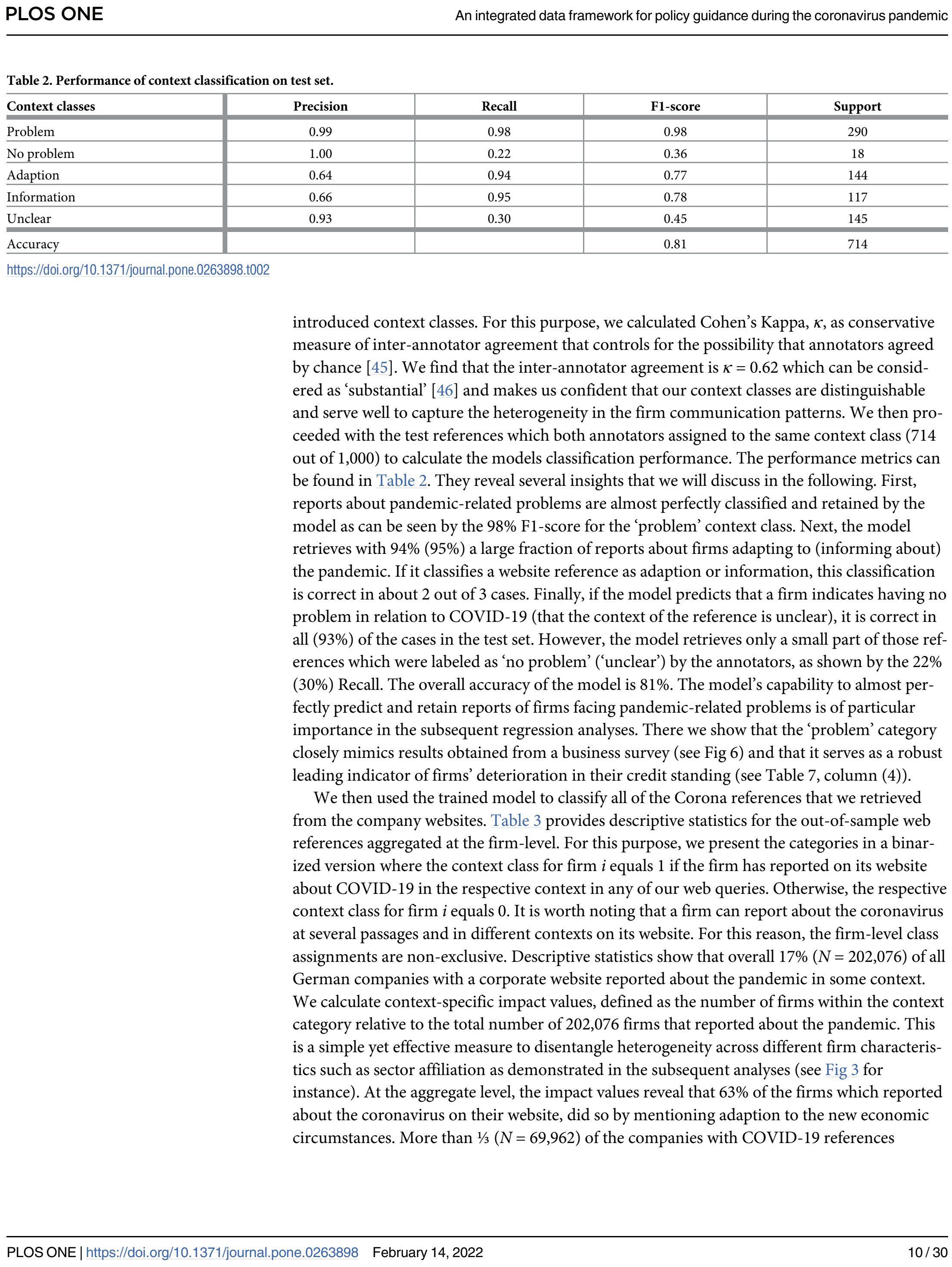


Table 2. Performance of context classification on test set.

Context classes	Precision	Recall	F1-score	Support
Problem	0.99	0.98	0.98	290
No problem	1.00	0.22	0.36	18
Adaption	0.64	0.94	0.77	144
Information	0.66	0.95	0.78	117
Unclear	0.93	0.30	0.45	145
Accuracy			0.81	714

https://doi.org/10.1371/journal.pone.0263898.t002

introduced context classes. For this purpose, we calculated Cohen's Kappa, κ, as conservative measure of inter-annotator agreement that controls for the possibility that annotators agreed by chance [45]. We find that the inter-annotator agreement is $\kappa = 0.62$ which can be considered as 'substantial' [46] and makes us confident that our context classes are distinguishable and serve well to capture the heterogeneity in the firm communication patterns. We then proceeded with the test references which both annotators assigned to the same context class (714 out of 1,000) to calculate the models classification performance. The performance metrics can be found in Table 2. They reveal several insights that we will discuss in the following. First, reports about pandemic-related problems are almost perfectly classified and retained by the model as can be seen by the 98% F1-score for the 'problem' context class. Next, the model retrieves with 94% (95%) a large fraction of reports about firms adapting to (informing about) the pandemic. If it classifies a website reference as adaption or information, this classification is correct in about 2 out of 3 cases. Finally, if the model predicts that a firm indicates having no problem in relation to COVID-19 (that the context of the reference is unclear), it is correct in all (93%) of the cases in the test set. However, the model retrieves only a small part of those references which were labeled as 'no problem' ('unclear') by the annotators, as shown by the 22% (30%) Recall. The overall accuracy of the model is 81%. The model's capability to almost perfectly predict and retain reports of firms facing pandemic-related problems is of particular importance in the subsequent regression analyses. There we show that the 'problem' category closely mimics results obtained from a business survey (see Fig 6) and that it serves as a robust leading indicator of firms' deterioration in their credit standing (see Table 7, column (4)).

We then used the trained model to classify all of the Corona references that we retrieved from the company websites. Table 3 provides descriptive statistics for the out-of-sample web references aggregated at the firm-level. For this purpose, we present the categories in a binarized version where the context class for firm *i* equals 1 if the firm has reported on its website about COVID-19 in the respective context in any of our web queries. Otherwise, the respective context class for firm *i* equals 0. It is worth noting that a firm can report about the coronavirus at several passages and in different contexts on its website. For this reason, the firm-level class assignments are non-exclusive. Descriptive statistics show that overall 17% (*N* = 202,076) of all German companies with a corporate website reported about the pandemic in some context. We calculate context-specific impact values, defined as the number of firms within the context category relative to the total number of 202,076 firms that reported about the pandemic. This is a simple yet effective measure to disentangle heterogeneity across different firm characteristics such as sector affiliation as demonstrated in the subsequent analyses (see Fig 3 for instance). At the aggregate level, the impact values reveal that 63% of the firms which reported about the coronavirus on their website, did so by mentioning adaption to the new economic circumstances. More than ⅓ (*N* = 69,962) of the companies with COVID-19 references

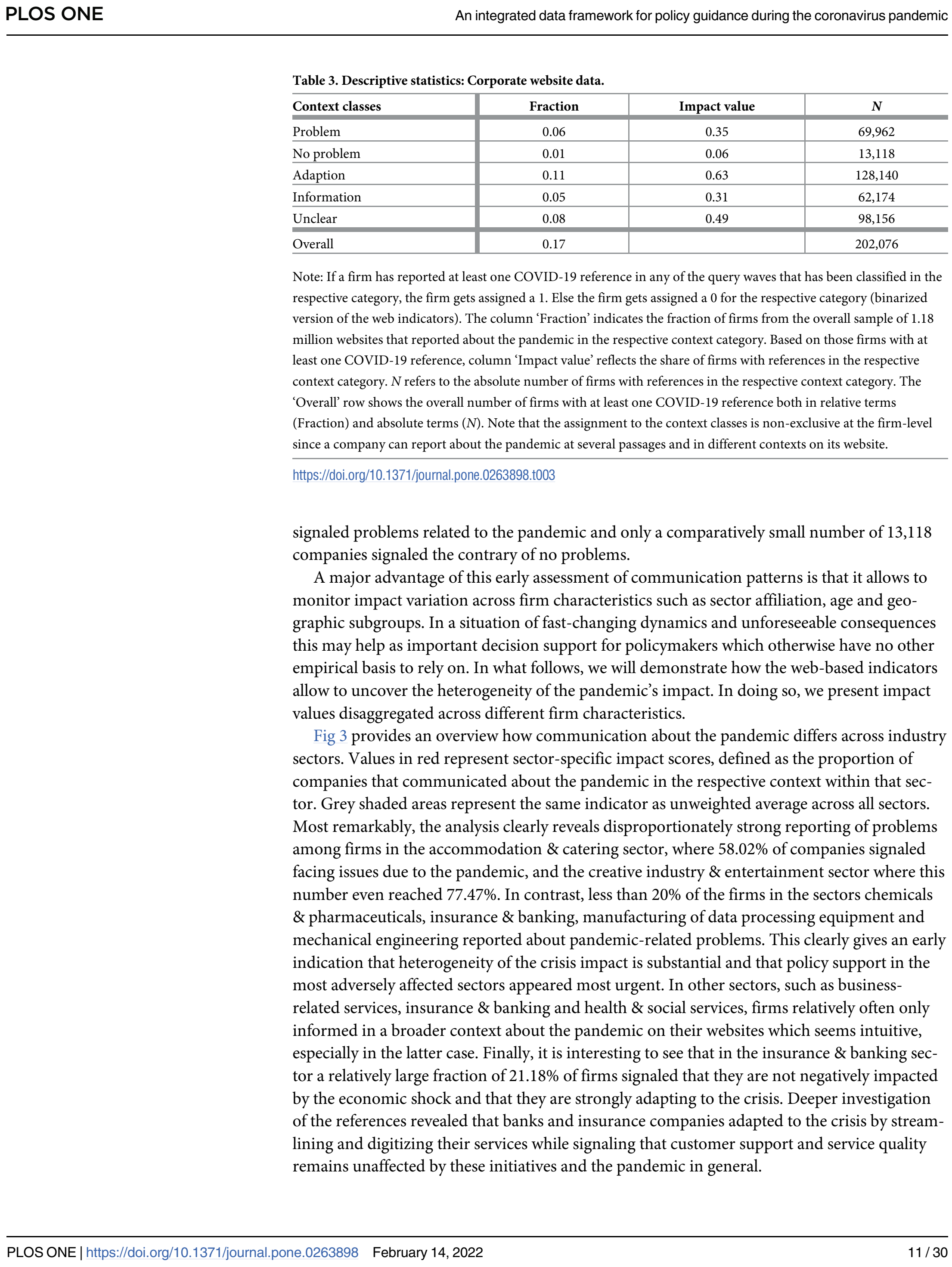


Table 3. Descriptive statistics: Corporate website data.

Context classes	Fraction	Impact value	*N*
Problem	0.06	0.35	69,962
No problem	0.01	0.06	13,118
Adaption	0.11	0.63	128,140
Information	0.05	0.31	62,174
Unclear	0.08	0.49	98,156
Overall	0.17		202,076

Note: If a firm has reported at least one COVID-19 reference in any of the query waves that has been classified in the respective category, the firm gets assigned a 1. Else the firm gets assigned a 0 for the respective category (binarized version of the web indicators). The column 'Fraction' indicates the fraction of firms from the overall sample of 1.18 million websites that reported about the pandemic in the respective context category. Based on those firms with at least one COVID-19 reference, column 'Impact value' reflects the share of firms with references in the respective context category. *N* refers to the absolute number of firms with references in the respective context category. The 'Overall' row shows the overall number of firms with at least one COVID-19 reference both in relative terms (Fraction) and absolute terms (*N*). Note that the assignment to the context classes is non-exclusive at the firm-level since a company can report about the pandemic at several passages and in different contexts on its website.

https://doi.org/10.1371/journal.pone.0263898.t003

signaled problems related to the pandemic and only a comparatively small number of 13,118 companies signaled the contrary of no problems.

A major advantage of this early assessment of communication patterns is that it allows to monitor impact variation across firm characteristics such as sector affiliation, age and geographic subgroups. In a situation of fast-changing dynamics and unforeseeable consequences this may help as important decision support for policymakers which otherwise have no other empirical basis to rely on. In what follows, we will demonstrate how the web-based indicators allow to uncover the heterogeneity of the pandemic's impact. In doing so, we present impact values disaggregated across different firm characteristics.

Fig 3 provides an overview how communication about the pandemic differs across industry sectors. Values in red represent sector-specific impact scores, defined as the proportion of companies that communicated about the pandemic in the respective context within that sector. Grey shaded areas represent the same indicator as unweighted average across all sectors. Most remarkably, the analysis clearly reveals disproportionately strong reporting of problems among firms in the accommodation & catering sector, where 58.02% of companies signaled facing issues due to the pandemic, and the creative industry & entertainment sector where this number even reached 77.47%. In contrast, less than 20% of the firms in the sectors chemicals & pharmaceuticals, insurance & banking, manufacturing of data processing equipment and mechanical engineering reported about pandemic-related problems. This clearly gives an early indication that heterogeneity of the crisis impact is substantial and that policy support in the most adversely affected sectors appeared most urgent. In other sectors, such as business-related services, insurance & banking and health & social services, firms relatively often only informed in a broader context about the pandemic on their websites which seems intuitive, especially in the latter case. Finally, it is interesting to see that in the insurance & banking sector a relatively large fraction of 21.18% of firms signaled that they are not negatively impacted by the economic shock and that they are strongly adapting to the crisis. Deeper investigation of the references revealed that banks and insurance companies adapted to the crisis by streamlining and digitizing their services while signaling that customer support and service quality remains unaffected by these initiatives and the pandemic in general.

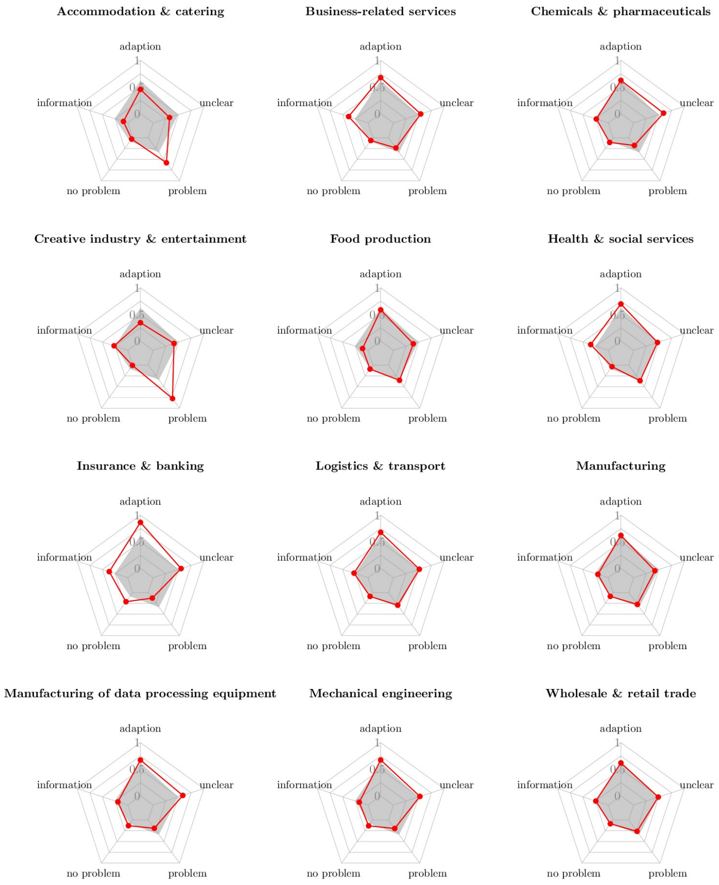


Fig 3. COVID-19 firm communication on corporate websites website-generated impact values at sector level. Note: Visualizations based on classified COVID-19 web references. If a firm reported at least one COVID-19 reference that has been classified in the respective context class in any of the web queries, the firm gets assigned a 1. Else the firm gets assigned a 0 for the respective class (binarized version of the web indicators). Red lines represent sector-specific impact values. Grey shaded areas represent unweighted average impact values across all sectors. Exact numerical values can be found in S6 Table.

https://doi.org/10.1371/journal.pone.0263898.g003

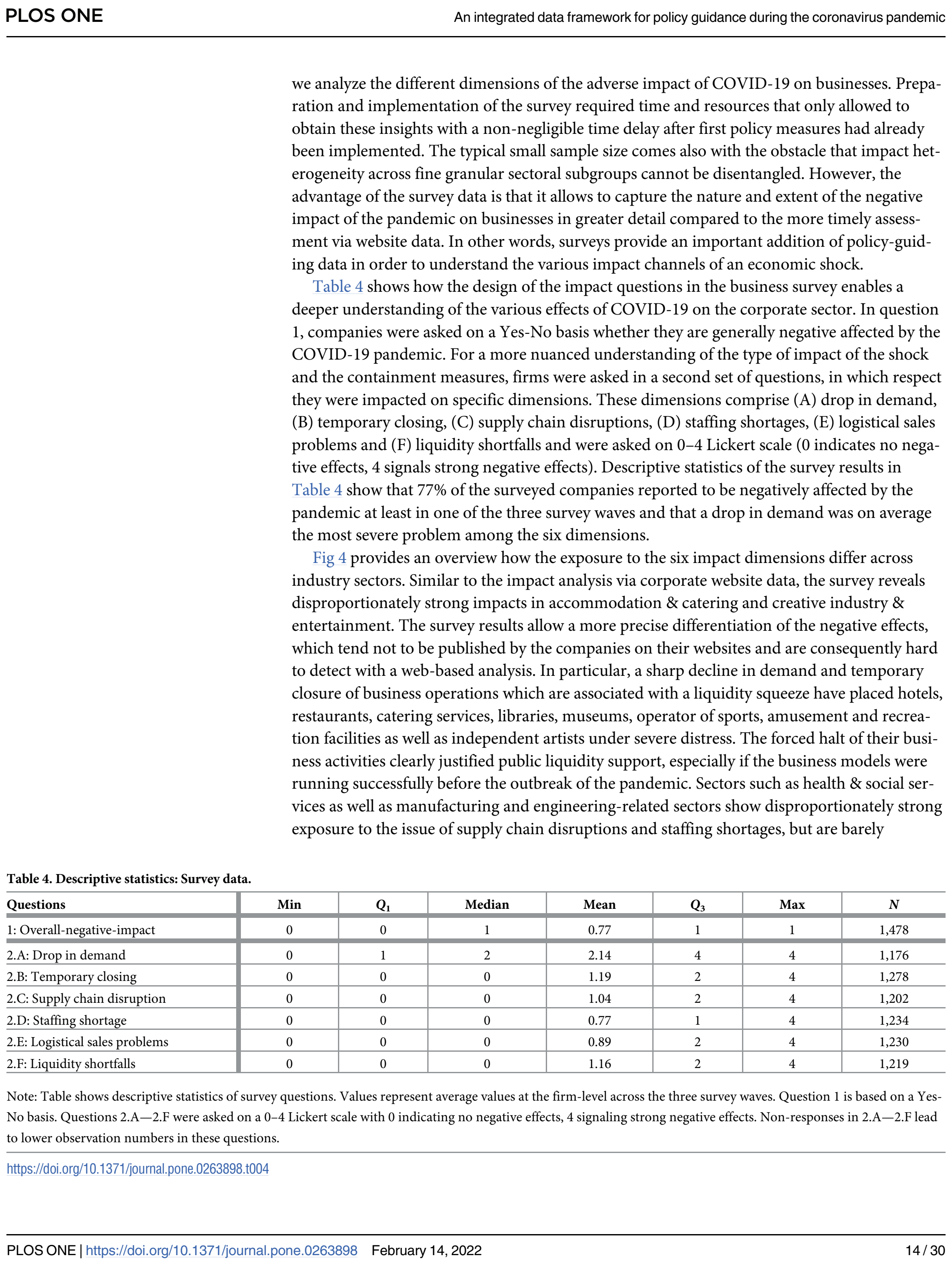
Given the large sample size of the webdata, these sectoral impacts can be further disaggregated at a finer level of granularity. To demonstrate this, we focus exemplary on the heterogeneity of firms reporting about problems *within* the wholesale sector. The wholesale sector is interesting for two reasons. First, the aggregated view on the trade sector, as displayed in the lower right corner in Fig 3, does only reveal that around 26% of the firms communicated about issues which is well below the (unweighted) average impact value of 34% across all sectors. Survey-based data usually do not allow to break this insight further down to a subsector level due to their relatively small sample size. Policymakers would thus be left with the information that negative impacts are below average for firms operating in trade, although some trade subsectors may be much more severely affected. Second, wholesale companies can provide important signals about the extent to which the national supply of certain goods and commodities may become tight. In the coronavirus pandemic, this has become a severe problem since international shutdown measures and changed consumer behavior have led to supply shortages of various goods. Problem reports in the wholesale sector disaggregated by single product markets, can help to signal the risk of supply issues at an early stage. While analyzing the relation between communication patterns and actual supply shortages is beyond the scope of this study, we still present the fine-granular impact values of distinct product markets within the wholesale sector. For this purpose, we follow the statistical classification of economic activities of the European Union [47] and assign all wholesale companies in our sample to close to 40 different product markets (see S7 Table for a detailed listing). From this disaggregation, we see that especially supply of household goods such as textiles but also manufacturing goods such as machineries, intermediates and related equipment were most adversely affected in the early stage of the pandemic. Basic supply such as food and beverages and, from a policy perspective also important, supply of pharmaceutical goods seemed to be less at stake since less than 15% of the respective wholesale companies reported about issues.

These near real-time insights into the heterogeneity of the Corona pandemic's impact on the business sector provide policymakers with a better understanding how early and more targeted impulse measures can be designed. Without the time to wait for official surveys to reveal the effects of the pandemic and shutdowns, we see this stage of our framework as an explorative analysis how governments can be assisted with empirical evidence from alternative data sources. While this section strongly focused on the heterogeneity across different industries and sectoral subgroups, it shall be clear that the presented approach can be used to set-up a monitoring system not only to track 'problem' sectors but also to unveil impact variation along further firm dimensions. For instance, the system can additionally account for regional differences based on the firms location (see, for example, [36] for an early version of the assessment of the Corona pandemic via corporate websites and S1 Fig).

3.2 Second stage: Follow-up survey-based effect differentiation

In a second stage, after firms have been exposed to the adverse economic environment for a critical period of time, we transfer our impact analysis from corporate website data to results obtained from a questionnaire-based survey which allows us to further differentiate the firm-level effects of the pandemic. The early insights from the first stage of our framework, serves here as valuable guide for the concrete design of the survey (e.g. formulation of questions, implementation of sampling strategy). For example, we could use the insights from the annotation and classification process in the first stage, to formulate relevant and specific survey questions concerning the types of problems companies were facing early on in the pandemic. This shows that the near real-time information obtained from the first stage of the framework allows for a more targeted design of follow-up surveys.

In order to guarantee a continuous update of policymakers information basis, the business survey has been conducted consecutively. Starting in mid of April, the survey has thus been repeated mid of June and end of September 2020. Over the course of the three survey waves, information on 1,478 distinct companies could be analyzed (the survey is a representative random sample of German companies, drawn from the MUP and stratified by firm size and industry affiliation—further details concerning sampling strategy and exact sample size for each of the survey waves can be found in S1 Appendix). Based on these consecutive surveys,



we analyze the different dimensions of the adverse impact of COVID-19 on businesses. Preparation and implementation of the survey required time and resources that only allowed to obtain these insights with a non-negligible time delay after first policy measures had already been implemented. The typical small sample size comes also with the obstacle that impact heterogeneity across fine granular sectoral subgroups cannot be disentangled. However, the advantage of the survey data is that it allows to capture the nature and extent of the negative impact of the pandemic on businesses in greater detail compared to the more timely assessment via website data. In other words, surveys provide an important addition of policy-guiding data in order to understand the various impact channels of an economic shock.

Table 4 shows how the design of the impact questions in the business survey enables a deeper understanding of the various effects of COVID-19 on the corporate sector. In question 1, companies were asked on a Yes-No basis whether they are generally negative affected by the COVID-19 pandemic. For a more nuanced understanding of the type of impact of the shock and the containment measures, firms were asked in a second set of questions, in which respect they were impacted on specific dimensions. These dimensions comprise (A) drop in demand, (B) temporary closing, (C) supply chain disruptions, (D) staffing shortages, (E) logistical sales problems and (F) liquidity shortfalls and were asked on 0–4 Lickert scale (0 indicates no negative effects, 4 signals strong negative effects). Descriptive statistics of the survey results in Table 4 show that 77% of the surveyed companies reported to be negatively affected by the pandemic at least in one of the three survey waves and that a drop in demand was on average the most severe problem among the six dimensions.

Fig 4 provides an overview how the exposure to the six impact dimensions differ across industry sectors. Similar to the impact analysis via corporate website data, the survey reveals disproportionately strong impacts in accommodation & catering and creative industry & entertainment. The survey results allow a more precise differentiation of the negative effects, which tend not to be published by the companies on their websites and are consequently hard to detect with a web-based analysis. In particular, a sharp decline in demand and temporary closure of business operations which are associated with a liquidity squeeze have placed hotels, restaurants, catering services, libraries, museums, operator of sports, amusement and recreation facilities as well as independent artists under severe distress. The forced halt of their business activities clearly justified public liquidity support, especially if the business models were running successfully before the outbreak of the pandemic. Sectors such as health & social services as well as manufacturing and engineering-related sectors show disproportionately strong exposure to the issue of supply chain disruptions and staffing shortages, but are barely

Table 4. Descriptive statistics: Survey data.

Questions	Min	Q_1	Median	Mean	Q_3	Max	*N*
1: Overall-negative-impact	0	0	1	0.77	1	1	1,478
2.A: Drop in demand	0	1	2	2.14	4	4	1,176
2.B: Temporary closing	0	0	0	1.19	2	4	1,278
2.C: Supply chain disruption	0	0	0	1.04	2	4	1,202
2.D: Staffing shortage	0	0	0	0.77	1	4	1,234
2.E: Logistical sales problems	0	0	0	0.89	2	4	1,230
2.F: Liquidity shortfalls	0	0	0	1.16	2	4	1,219

Note: Table shows descriptive statistics of survey questions. Values represent average values at the firm-level across the three survey waves. Question 1 is based on a Yes-No basis. Questions 2.A—2.F were asked on a 0–4 Lickert scale with 0 indicating no negative effects, 4 signaling strong negative effects. Non-responses in 2.A—2.F lead to lower observation numbers in these questions.

https://doi.org/10.1371/journal.pone.0263898.t004

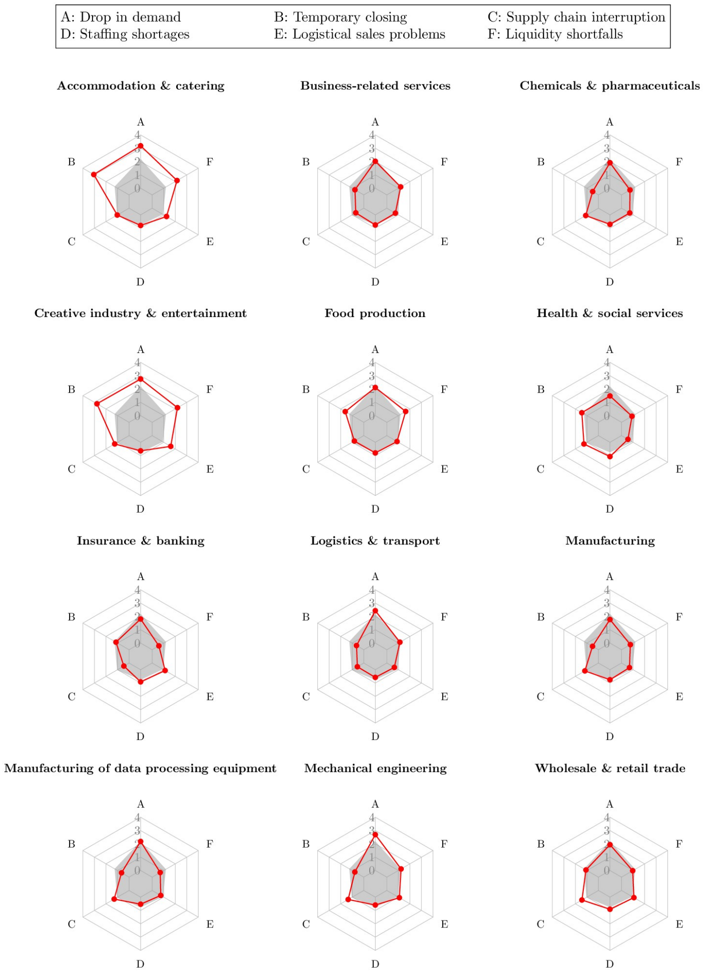


Fig 4. COVID-19 firm exposure at sector level based on survey results. Note: Visualizations based on survey questions A—F which were asked on a 0–4 Lickert scale with 0 indicating no negative effects, 4 signaling strong negative effects. Red lines represent sector-specific impact values. Grey shaded areas represent unweighted average impact values across all sectors. All values are averaged at the firm-level across the three survey waves.

https://doi.org/10.1371/journal.pone.0263898.g004

confronted with declining demand numbers and liquidity shocks. It is clear that for firms in these sectors, the priority of policy should not be to provide liquidity support. An important issue for these firms is maintaining relationships with their stakeholders. Building these relationships is costly, and maintaining them despite the economic downturn is key to a successful recovery for many of these firms. Therefore, policymakers are challenged to renew and reevaluate their policy toolkit to find new tools to help companies maintain their stakeholder relationships with workers and suppliers during economic downturns [9].

The second 'follow-up' stage of our proposed framework clearly demonstrated that, based on survey data, businesses in the accommodation, arts, and entertainment sectors have been facing strong liquidity bottlenecks, which in light of often unchanged fixed cost obligations poses a high risk of financial insolvency. In the third 'retrospective' stage of our framework, we more closely focus on this liquidation risk by analyzing the change in corporate solvency information in response to the COVID-19 crisis.

3.3 Third stage: Retrospective liquidation risk analysis

A major economic threat of COVID-19 has been and, given the possibility of recurring shutdown measures, continues to be, is the risk that firms with sound business models and decent financial performance before the outbreak of the pandemic are forced into insolvency. For economic policymakers it is important to understand if and where in the economy firms are at risk to leave the market permanently. Depending on size and strategic importance of impacted industries, this could imply high costs in terms of losses in jobs and output. In a third and last stage of our framework, we thus focus on this liquidation risk by transferring our impact analysis from corporate websites (first stage) and survey data (second stage) to firm-specific credit rating information which gives a much conciser picture to what extent the pandemic has materialized in the firms' financial solvency. For this purpose, we examine credit rating updates in the crisis period for more than 870,000 German companies. While firm-specific credit rating data reflect very precise information concerning the firm's financial standing and in case of substantial credit rating downgrades signals risk of financial insolvency [48, 49], the reassessment of firms' solvency by rating agencies is time and resource expensive. Generally, we find that on average the time between two credit evaluations equals 18 months. Typically, if credit information of a firm is requested more often by an external creditor, a company will be re-evaluated more frequently. However, the rating capacity is largely tied to the headcount limitations of the rating agency. For this reason, only after a certain time a critical mass of rating updates becomes available to infer the heterogeneity of the crisis effects on companies' solvency.

The credit rating data that we analyze in the third stage of our framework is generated by Creditreform, Germany's leading credit agency. Creditreform assesses the creditworthiness of the near universe of active companies in Germany. The credit rating information is included for close to all firms in the MUP which allows to merge the ratings with the corporate website data (this becomes relevant in Section 4 of this study where we show how the webdata serves as leading indicator for later credit rating movements). Creditreform's corporate solvency index is based on a rich information set that closely mirrors a company's financial situation. Creditreform regularly investigates, among other things, information on the firm's payment discipline, its legal structure, credit evaluations of banks, caps in credit lines and further risk indicators based on the firm's financial accounts and incorporates this set of information into its rating score [50]. Different weights are attached to these metrics according to their importance for determining a firm's risk of defaulting on a loan. Overall, the rating index ranges from 100 to 500 with a higher index signaling a worse financial standing [51]. It is worth mentioning that the credit rating index suffers a discontinuity as, in case of a 'insufficient' creditworthiness, it takes on a value of 600. We truncate credit ratings of 600 to a value 500—the worst possible rating in our analysis. We do so since our main variable of interest is the *update* in the rating index which can only be reasonably calculated if the index has continuous support.

Table 5. Descriptive statistics: Credit rating data.

Variables	Min	Q_1	Median	Mean	Q_3	Max
Δr_t	-315	-3	0	3.3	0	357
Date of update	2-Jun-20	3-Sep-20	30-Oct-20	21-Oct-20	8-Dec-20	9-Apr-21

Note: Table shows descriptive statistics of the rating updates and statistics of the dates of the rating revaluations. $\Delta r_t = 0$ means that the revaluation of the company has not led to any changes in its solvency compared to the pre-crisis period. Q_1 refers to the first quartile and Q_3 to the third quartile, respectively. The distribution of the dates of rating updates shows that more than 75% of the updates took place in 2020 and the latest update in the sample was conducted beginning of April 2021.

https://doi.org/10.1371/journal.pone.0263898.t005

The level of the rating itself is little informative for inferring the effects of the COVID-19 crisis on the corporate sector. The change in the firms' credit rating, Δr_t, in contrast, precisely reflects to what extent a company has been down- or upgraded after the shock has hit the German economy. For that purpose, we consider all credit rating updates that have been conducted by Creditreform after June 1, 2020. We choose this date as it ensures that sufficient time has passed since the onset of the crisis to reflect COVID-related effects in the rating updates. The update in a firm's credit rating is defined as simple difference between the new rating index and the index before the update (i.e. before COVID-19) with a positive value indicating a downgrade and a negative value signaling an upgrade.

$$\Delta r_t = r_t - r_{t-x} \tag{2}$$

Reassessments of the rating are conducted irregularly such that the time between two updates, x, varies. On average, the time between two updates in our sample equals 18 months.

Descriptive statistics in Table 5 show that most of the distribution is centered around 0, implying that a substantial number of firms experienced only minor changes in their credit ratings during the COVID-19 crisis. However, taking a closer look at the distribution of the rating updates across industry sectors in Fig 5 reveals an interesting pattern: sectors that, according to our first and second stage results, are severely affected such as logistics & transport, accommodation & catering and creative industry & entertainment but also supposedly winners of the crisis, most notably health & social services, follow a bimodal distribution. Comparing the crisis distribution with the pre-crisis distribution (indicated as dashed green line) suggests that this bimodality is indeed the result of the COVID-19 crisis. This means that major rating downgrades and upgrades are more likely in times of crisis than in normal times when a sector is severely affected by the crisis. We see this as strong hint that the pandemic shock has strongly materialized in severely affected firms' financial solvency with strong differences across sectors.

Moreover, the minimum and maximum values of Δr_t in Table 5 show that there are some companies that have experienced large downgrades or upgrades in their credit ratings. To shed more light on this occurrence, Table 6 shows the fraction of firms with a substantial rating downgrade of more than 50 index points within the respective sector. We see again that logistics & transport, accommodation & catering and creative industry & entertainment show a relatively high fraction of firms which experienced a substantial downgrade in their ratings compared to less affected industries as well as compared to pre-crisis numbers. These high fractions of substantial rating downgrades reflect a relatively high insolvency risk in the respective industries. Despite the substantial policy support that these sectors received, this hints to a non-negligible number of market exits if support measures will cease before the firms have overcome the financial repercussions of the shock.

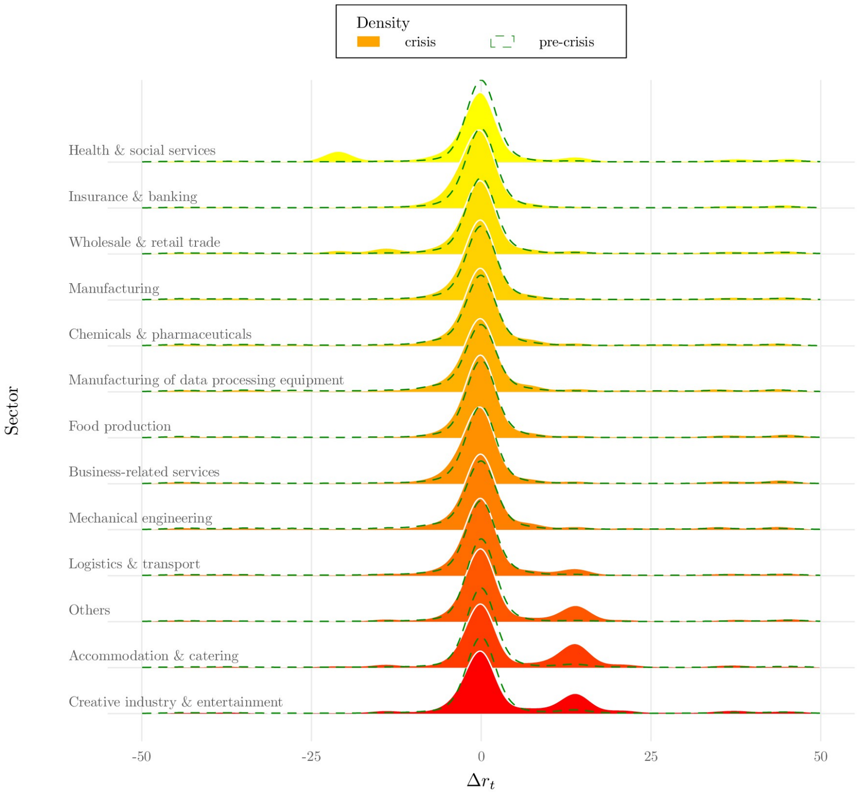


Fig 5. COVID-19 effects on corporate solvency at sector level. Note: Figure shows distribution of credit rating updates both during COVID-19 (yellow to red palette) and before COVID-19 (dashed green line). Densities are based on a Gaussian smoothing kernel with a bandwidth of 2.

https://doi.org/10.1371/journal.pone.0263898.g005

In the last stage of our framework, we have focused on the structural risk of firms being forced to leave the market. Economic shutdown measures and drop in consumer demand have made this a particular concern of policymakers over the first months of the pandemic. Clearly, concerns about systematic bankruptcies have been specific to the COVID-19 pandemic and it is likely that the nature of future shocks will cause decision-makers to focus on different outcome variables. In the following section, we thus outline how our proposed framework can be extended to become applicable to a wider set of economic shocks and thus a useful tool to provide economic decision-makers with timely insights.

3.4 Generalizability to other types of shocks

While the approach presented to track the early impact of an economic shock using corporate communication patterns is specific to the coronavirus pandemic, the idea to integrate timely online sources with more traditional but less timely policy data can also be useful in other crisis

Table 6. Distribution of extreme rating downgrades.

Sector	crisis		pre-crisis	
	N	Substantial downgrades in %	*N*	Substantial downgrades in %
Insurance & banking	34,768	3.0	35,087	1.7
Manufacturing of data processing equipment	4,512	3.1	4,406	2.7
Chemicals & pharmaceuticals	7,204	3.1	7,000	2.4
Manufacturing	224,813	3.5	204,613	2.7
Food production	10,420	3.8	10,311	2.9
Health & social services	62,633	4.0	57,466	2.0
Mechanical engineering	12,254	4.1	12,361	2.8
Business-related services	227,957	4.3	232,576	2.4
Others	14,259	4.8	12,511	2.0
Wholesale & retail trade	173,619	4.8	169,109	2.9
Logistics & transport	39,164	5.9	37,817	3.7
Creative industry & entertainment	13,865	8.7	12,967	3.9
Accommodation & catering	44,692	9.0	36,289	4.6
Overall	870,195	4.5	832,513	2.7

Note: Table shows fraction of firms with major credit rating downgrades by industry sector in percent. Substantial downgrades are defined as credit rating downgrades of more than 50 index points ($\Delta r_t > 50$). Pre-crisis numbers refer to the year 2018.

https://doi.org/10.1371/journal.pone.0263898.t006

scenarios. Timely impact data from real-time online sources may not only reveal early heterogeneity at granular level or serve as leading indicators (as we will show in Section 4), but are equally important to design targeted surveys which in turn reveal more granular information how firms are affected by a shock and how and why possible heterogeneity in these effects may translate into heterogeneous firm outcomes. Integrating these different sources of information into a common framework allows policymakers not only to react more swiftly and targeted, but also allows to design medium to long-term stimulus packages based on a rich set of information that has been continuously updated over all stages of the shock. For example, one could think of immediate subsidies only for firms in hard-hit sectors (as evidenced by, e.g., a real-time online source) which suffered temporary liquidity constraints but are characterized by a robust pre-crisis performance (as indicated by, e.g., credit rating agencies and public annual reports). Moreover, policy decisions can be justified more easily if they are backed by empirical evidence. In this sense, the integration of real-time data sources for policy guidance is an important step towards evidence-based decision-making if unexpected dynamics require fast action. We suggest that this holistic approach to policy guidance, by combining different sources of information, bears the potential to be applicable to a wider range of economic shocks. As demonstrated in this paper, this requires a strategy of complementing specific crisis-related data sources. From early stage insights that are generated from real-time sources, to a follow-up stage based on targeted surveys, to a retrospective stage in the aftermath of the shock focusing on relevant outcome variables that have been identified in the earlier stages.

Ideally, a universally applicable system would be available to political decision-makers for this purpose in the future. However, a framework that can be used universally and on short notice requires that a large part of the early stage analysis is automated to a greater extent than in our study. This applies in particular to the selection of meaningful keywords, but also to the definition of 'impact' classes. Both could be done in the future, for example, by monitoring news streams. Business news articles could be used to filter relevant topics based on their popularity and, for instance, a sentiment analysis. Relevant keywords could then be extracted automatically from all of these articles (based on unsupervised learning such as topic modeling whose topic-specific probability vectors over the vocabulary allow the identification of relevant keywords). These keywords would then be the input for a keyword search as described in our article, which would then be used as the basis for constructing classes (e.g., via clustering). Another promising approach for an extensive automation of our proposed framework could be the so-called zero shot classification. Zero shot does not require text analysis models to be fine-tuned for specific classifications, as is common in transfer learning. Instead, one relies on the general text understanding of the model learned in the original pre-training and works by posing each candidate label as a 'hypothesis' and the text sequence which we want to classify as the 'premise'. The zero shot model then estimates whether the hypothesis and premise match or not, respectively whether the assumption formulated in the hypothesis is confirmed by the text. Thus, without the time-consuming manual labeling of training data (as we have done in this paper), one can directly ask content-related questions with regard to individual sentences or text passages and, in the best case, receive a reliable answer.

Form this line of argumentation, it becomes apparent that more work should be dedicated in researching how real-time data sources can complement traditional forms of policy data, especially if empirical evidence is required for immediate government response. However, even if real-time data can be made accessible for policy guidance, an important question remains: What value do these information sources carry? In the next section, we show that the near real-time assessment via company communication patterns closely resembles heterogeneous effect estimates across various firm characteristics generated from survey responses. Moreover, we demonstrate that the webdata-generated impact values serve as leading indicators for companies' credit rating movements.

4 Assessing the predictive quality of early stage web-based impact indicators

The previous section has shown that all of the proposed data sources—corporate website data, survey data and credit rating data—hint to a strong degree of heterogeneity across economic sectors. While survey and credit rating data only revealed such patterns with a non-negligible time delay after the economic shock, corporate communication patterns retrieved from company websites indicated this heterogeneity at near real-time. A central question is to what extent the generated web indicators have predictive power in capturing the actual medium-term effects of the coronavirus shock. Clearly, predictive power is an important prerequisite for the web indicators to be useful for policymakers. Only if the webdata's early indication generates reliable insights, it bears the potential to help policymakers tailor their response measures and effectively channel economic assistance where it is needed most.

We assess the predictive value of the early web indicators by two distinct analyses: First, we compare the relationship between several firm characteristics and the negative shock exposure based on two identical regression specifications. The only difference between the two regressions is that we exchange the target variable, which in the first regression is generated from company website information (data from the first 'ad hoc' stage), while in the second regression it stems from the business survey (data from the second 'follow-up' stage). Second, based on a sub-sample of firms for which we have both COVID-19 web references as well as credit rating updates, we analyze to what extent the classified web references serve as leading indicators for later changes in the firms' credit rating.

To examine the statistical relationships between various firm characteristics and the negative effects of the COVID-19 shock on firms, we conduct a Probit regression. More precisely, we regress a binary negative impact variable on age, size and sector characteristics.

$$Y_{k,i} = \alpha + \boldsymbol{\beta} \boldsymbol{A}_i + \boldsymbol{\gamma} \boldsymbol{S}_i + \boldsymbol{\delta} \boldsymbol{I}_i + \epsilon_i \tag{3}$$

with

$$Y_{k,i} = \begin{cases} \text{problem}_i, & \text{if } k = \text{Webdata} \\ \text{overall} - \text{negative} - \text{impact}_i, & \text{if } k = \text{Survey} \end{cases}$$

and ***A***, ***S*** and ***I*** matrices of company age, size and sector controls, respectively.

First, we conduct the regression estimation based on the corporate website observations. The dependent negative impact variable, $Y_{Webdata,i}$ equals 1 if the firm has reported a problem on its website and 0 otherwise. Second, we estimate the same regression specification based on the survey observations where the dependent variable reflects the first question in the survey: 'Has the coronavirus pandemic had negative economic effects on your company so far?' If the firm confirmed the question, $Y_{Survey,i}$ equals 1, otherwise it is 0.

Fig 6 visualizes the estimation results of both Probit regressions. Effect estimates need to be interpreted relative to the reference firm which is defined as an incumbent (10 years and older), micro company (less than 10 employees) in the accommodation and catering sector.

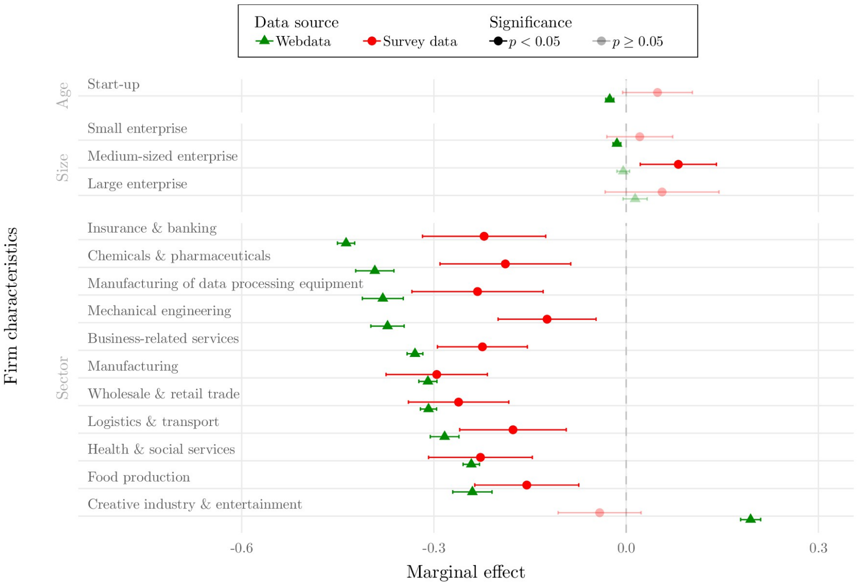
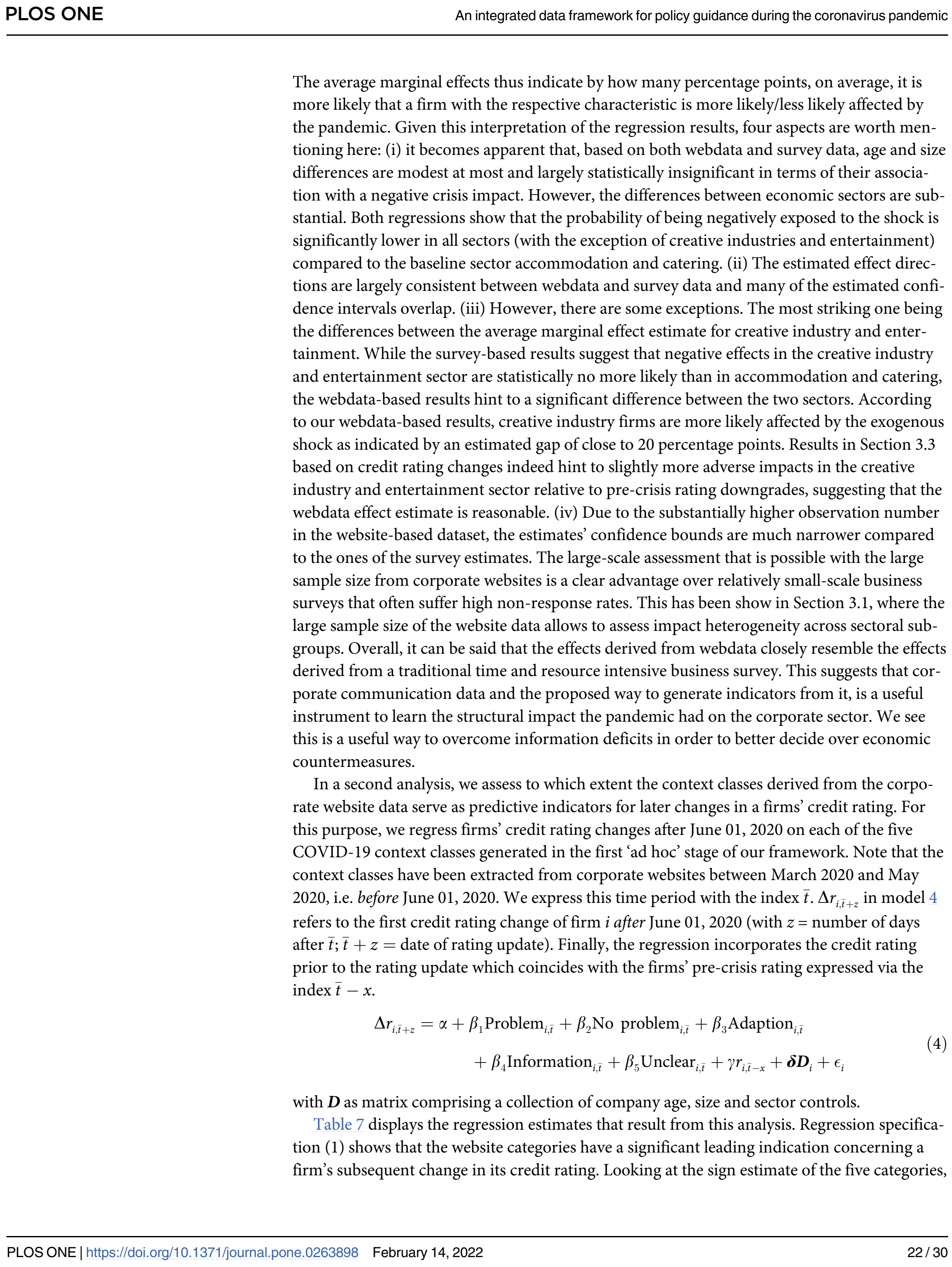


Fig 6. Comparison webdata and survey data effect estimates. Note: Figure shows average marginal effect estimates and corresponding 95%-confidence intervals of model 3 where the dependent variable (negative impact) is generated from webdata (green) and survey data (red). Dependent variable from webdata reflects whether the firm has reported a 'problem' reference on its corporate website in any of the web queries. Dependent variable from survey data refers to the question whether the firm has suffered negative impacts due to the pandemic in any of the three survey waves. Shaded estimates signal statistically insignificant effects at the 5% level. Incumbent firms (10 years and older) serve as baseline age group, micro-enterprises (number of employees ≤ 10) as baseline size group, accommodation and catering serves as baseline sector among the sector dummies. Marginal effects need to be interpreted relative to the baseline group(s).

https://doi.org/10.1371/journal.pone.0263898.g006



The average marginal effects thus indicate by how many percentage points, on average, it is more likely that a firm with the respective characteristic is more likely/less likely affected by the pandemic. Given this interpretation of the regression results, four aspects are worth mentioning here: (i) it becomes apparent that, based on both webdata and survey data, age and size differences are modest at most and largely statistically insignificant in terms of their association with a negative crisis impact. However, the differences between economic sectors are substantial. Both regressions show that the probability of being negatively exposed to the shock is significantly lower in all sectors (with the exception of creative industries and entertainment) compared to the baseline sector accommodation and catering. (ii) The estimated effect directions are largely consistent between webdata and survey data and many of the estimated confidence intervals overlap. (iii) However, there are some exceptions. The most striking one being the differences between the average marginal effect estimate for creative industry and entertainment. While the survey-based results suggest that negative effects in the creative industry and entertainment sector are statistically no more likely than in accommodation and catering, the webdata-based results hint to a significant difference between the two sectors. According to our webdata-based results, creative industry firms are more likely affected by the exogenous shock as indicated by an estimated gap of close to 20 percentage points. Results in Section 3.3 based on credit rating changes indeed hint to slightly more adverse impacts in the creative industry and entertainment sector relative to pre-crisis rating downgrades, suggesting that the webdata effect estimate is reasonable. (iv) Due to the substantially higher observation number in the website-based dataset, the estimates' confidence bounds are much narrower compared to the ones of the survey estimates. The large-scale assessment that is possible with the large sample size from corporate websites is a clear advantage over relatively small-scale business surveys that often suffer high non-response rates. This has been show in Section 3.1, where the large sample size of the website data allows to assess impact heterogeneity across sectoral subgroups. Overall, it can be said that the effects derived from webdata closely resemble the effects derived from a traditional time and resource intensive business survey. This suggests that corporate communication data and the proposed way to generate indicators from it, is a useful instrument to learn the structural impact the pandemic had on the corporate sector. We see this is a useful way to overcome information deficits in order to better decide over economic countermeasures.

In a second analysis, we assess to which extent the context classes derived from the corporate website data serve as predictive indicators for later changes in a firms' credit rating. For this purpose, we regress firms' credit rating changes after June 01, 2020 on each of the five COVID-19 context classes generated in the first 'ad hoc' stage of our framework. Note that the context classes have been extracted from corporate websites between March 2020 and May 2020, i.e. *before* June 01, 2020. We express this time period with the index $\bar{t}$. $\Delta r_{i,\bar{t}+z}$ in model 4 refers to the first credit rating change of firm *i after* June 01, 2020 (with z = number of days after $\bar{t}$; $\bar{t}+z=$ date of rating update). Finally, the regression incorporates the credit rating prior to the rating update which coincides with the firms' pre-crisis rating expressed via the index $\bar{t}-x$.

$$\begin{aligned}\Delta r_{i,\bar{t}+z} = \alpha + \beta_1 \text{Problem}_{i,\bar{t}} + \beta_2 \text{No problem}_{i,\bar{t}} + \beta_3 \text{Adaption}_{i,\bar{t}} \\ + \beta_4 \text{Information}_{i,\bar{t}} + \beta_5 \text{Unclear}_{i,\bar{t}} + \gamma r_{i,\bar{t}-x} + \boldsymbol{\delta D}_i + \epsilon_i\end{aligned} \qquad (4)$$

with $\boldsymbol{D}$ as matrix comprising a collection of company age, size and sector controls.

Table 7 displays the regression estimates that result from this analysis. Regression specification (1) shows that the website categories have a significant leading indication concerning a firm's subsequent change in its credit rating. Looking at the sign estimate of the five categories,

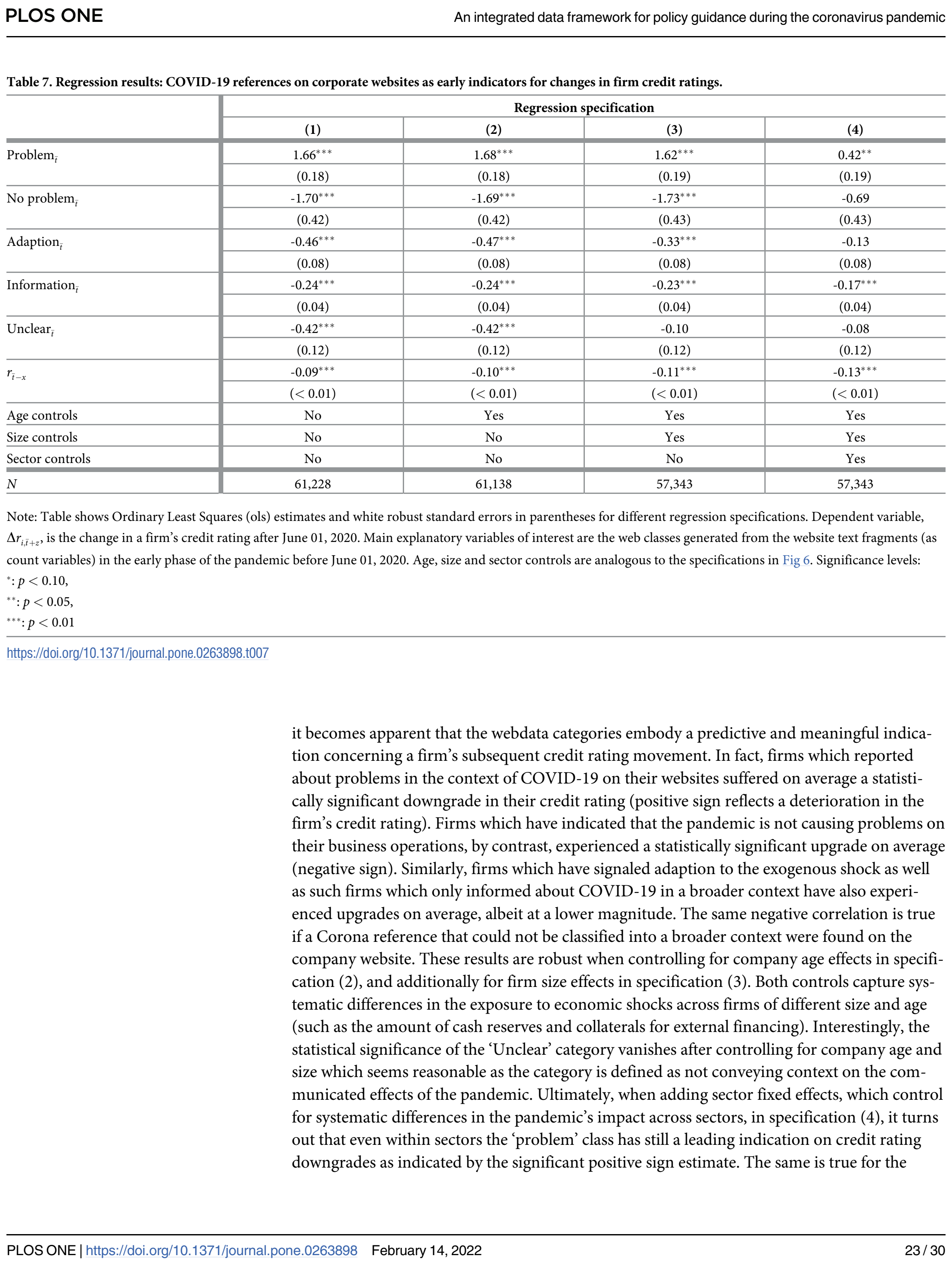


Table 7. Regression results: COVID-19 references on corporate websites as early indicators for changes in firm credit ratings.

	Regression specification			
	(1)	(2)	(3)	(4)
Problem$_i$	1.66***	1.68***	1.62***	0.42**
	(0.18)	(0.18)	(0.19)	(0.19)
No problem$_i$	-1.70***	-1.69***	-1.73***	-0.69
	(0.42)	(0.42)	(0.43)	(0.43)
Adaption$_i$	-0.46***	-0.47***	-0.33***	-0.13
	(0.08)	(0.08)	(0.08)	(0.08)
Information$_i$	-0.24***	-0.24***	-0.23***	-0.17***
	(0.04)	(0.04)	(0.04)	(0.04)
Unclear$_i$	-0.42***	-0.42***	-0.10	-0.08
	(0.12)	(0.12)	(0.12)	(0.12)
r_{t-x}	-0.09***	-0.10***	-0.11***	-0.13***
	($<$ 0.01)	($<$ 0.01)	($<$ 0.01)	($<$ 0.01)
Age controls	No	Yes	Yes	Yes
Size controls	No	No	Yes	Yes
Sector controls	No	No	No	Yes
N	61,228	61,138	57,343	57,343

Note: Table shows Ordinary Least Squares (ols) estimates and white robust standard errors in parentheses for different regression specifications. Dependent variable, $\Delta r_{i,t+z}$, is the change in a firm's credit rating after June 01, 2020. Main explanatory variables of interest are the web classes generated from the website text fragments (as count variables) in the early phase of the pandemic before June 01, 2020. Age, size and sector controls are analogous to the specifications in Fig 6. Significance levels:

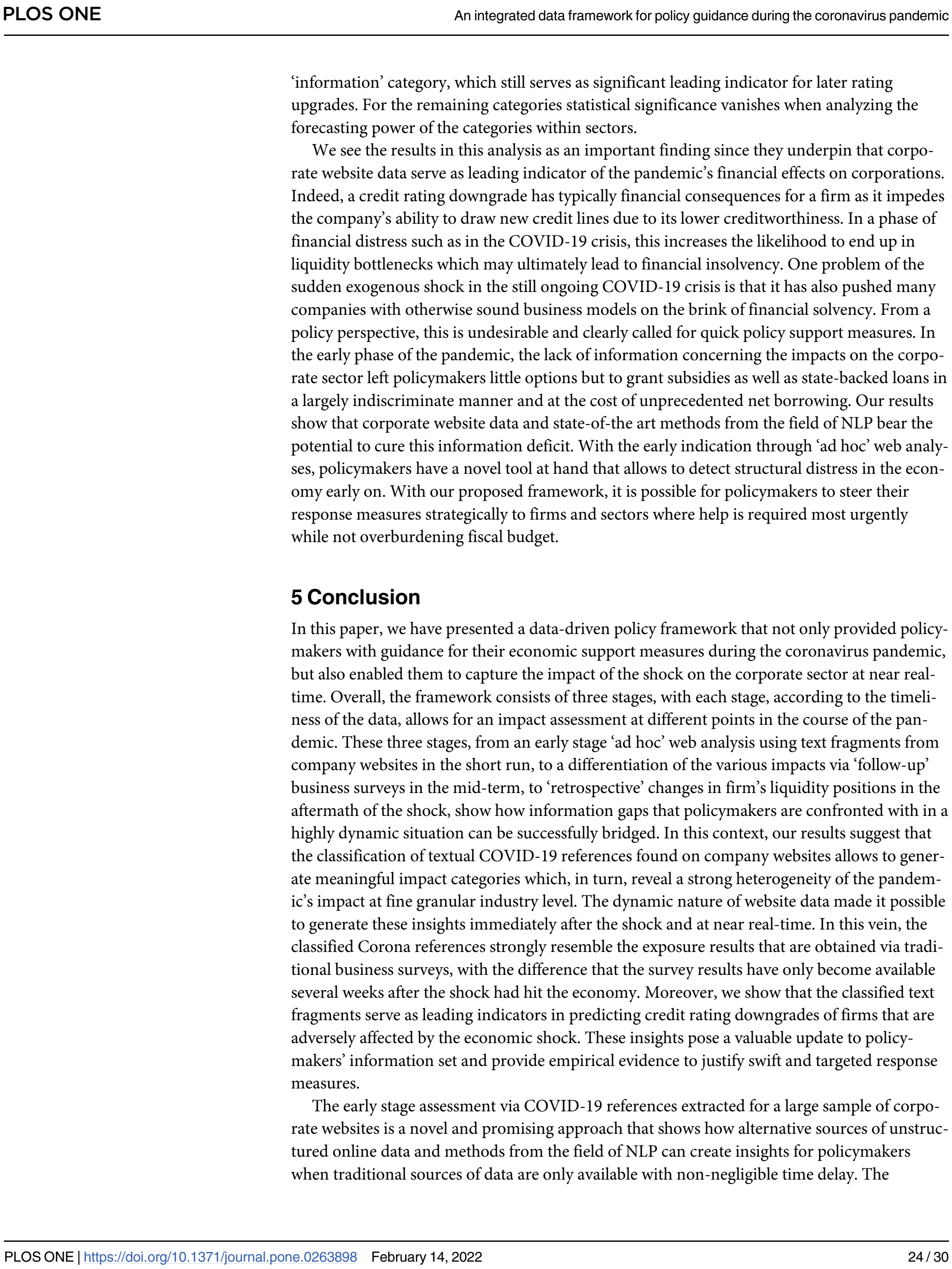
*: $p < 0.10$,

**: $p < 0.05$,

***: $p < 0.01$

https://doi.org/10.1371/journal.pone.0263898.t007

it becomes apparent that the webdata categories embody a predictive and meaningful indication concerning a firm's subsequent credit rating movement. In fact, firms which reported about problems in the context of COVID-19 on their websites suffered on average a statistically significant downgrade in their credit rating (positive sign reflects a deterioration in the firm's credit rating). Firms which have indicated that the pandemic is not causing problems on their business operations, by contrast, experienced a statistically significant upgrade on average (negative sign). Similarly, firms which have signaled adaption to the exogenous shock as well as such firms which only informed about COVID-19 in a broader context have also experienced upgrades on average, albeit at a lower magnitude. The same negative correlation is true if a Corona reference that could not be classified into a broader context were found on the company website. These results are robust when controlling for company age effects in specification (2), and additionally for firm size effects in specification (3). Both controls capture systematic differences in the exposure to economic shocks across firms of different size and age (such as the amount of cash reserves and collaterals for external financing). Interestingly, the statistical significance of the 'Unclear' category vanishes after controlling for company age and size which seems reasonable as the category is defined as not conveying context on the communicated effects of the pandemic. Ultimately, when adding sector fixed effects, which control for systematic differences in the pandemic's impact across sectors, in specification (4), it turns out that even within sectors the 'problem' class has still a leading indication on credit rating downgrades as indicated by the significant positive sign estimate. The same is true for the



'information' category, which still serves as significant leading indicator for later rating upgrades. For the remaining categories statistical significance vanishes when analyzing the forecasting power of the categories within sectors.

We see the results in this analysis as an important finding since they underpin that corporate website data serve as leading indicator of the pandemic's financial effects on corporations. Indeed, a credit rating downgrade has typically financial consequences for a firm as it impedes the company's ability to draw new credit lines due to its lower creditworthiness. In a phase of financial distress such as in the COVID-19 crisis, this increases the likelihood to end up in liquidity bottlenecks which may ultimately lead to financial insolvency. One problem of the sudden exogenous shock in the still ongoing COVID-19 crisis is that it has also pushed many companies with otherwise sound business models on the brink of financial solvency. From a policy perspective, this is undesirable and clearly called for quick policy support measures. In the early phase of the pandemic, the lack of information concerning the impacts on the corporate sector left policymakers little options but to grant subsidies as well as state-backed loans in a largely indiscriminate manner and at the cost of unprecedented net borrowing. Our results show that corporate website data and state-of-the art methods from the field of NLP bear the potential to cure this information deficit. With the early indication through 'ad hoc' web analyses, policymakers have a novel tool at hand that allows to detect structural distress in the economy early on. With our proposed framework, it is possible for policymakers to steer their response measures strategically to firms and sectors where help is required most urgently while not overburdening fiscal budget.

5 Conclusion

In this paper, we have presented a data-driven policy framework that not only provided policymakers with guidance for their economic support measures during the coronavirus pandemic, but also enabled them to capture the impact of the shock on the corporate sector at near real-time. Overall, the framework consists of three stages, with each stage, according to the timeliness of the data, allows for an impact assessment at different points in the course of the pandemic. These three stages, from an early stage 'ad hoc' web analysis using text fragments from company websites in the short run, to a differentiation of the various impacts via 'follow-up' business surveys in the mid-term, to 'retrospective' changes in firm's liquidity positions in the aftermath of the shock, show how information gaps that policymakers are confronted with in a highly dynamic situation can be successfully bridged. In this context, our results suggest that the classification of textual COVID-19 references found on company websites allows to generate meaningful impact categories which, in turn, reveal a strong heterogeneity of the pandemic's impact at fine granular industry level. The dynamic nature of website data made it possible to generate these insights immediately after the shock and at near real-time. In this vein, the classified Corona references strongly resemble the exposure results that are obtained via traditional business surveys, with the difference that the survey results have only become available several weeks after the shock had hit the economy. Moreover, we show that the classified text fragments serve as leading indicators in predicting credit rating downgrades of firms that are adversely affected by the economic shock. These insights pose a valuable update to policymakers' information set and provide empirical evidence to justify swift and targeted response measures.

The early stage assessment via COVID-19 references extracted for a large sample of corporate websites is a novel and promising approach that shows how alternative sources of unstructured online data and methods from the field of NLP can create insights for policymakers when traditional sources of data are only available with non-negligible time delay. The coronavirus pandemic has shown that in situations where policymakers need to respond quickly, but information deficits make it barely possible to determine where government assistance is channeled most efficiently, public aid measures are largely granted on a lump-sum basis. In fact, in Germany, this information deficit has led the Ministry of Finance to choose the 'bazooka' [52] instead of well-dosed and targeted liquidity injections as instrument to support companies in the early phase of the pandemic. Our framework is designed to help overcome information deficits that lead to otherwise undifferentiated support measures. In this context, we see this study as a first step towards real-time decision support for economic policymakers. Given further research and development, we argue that our framework can serve as a monitoring framework applicable to a wider range of economic shocks.

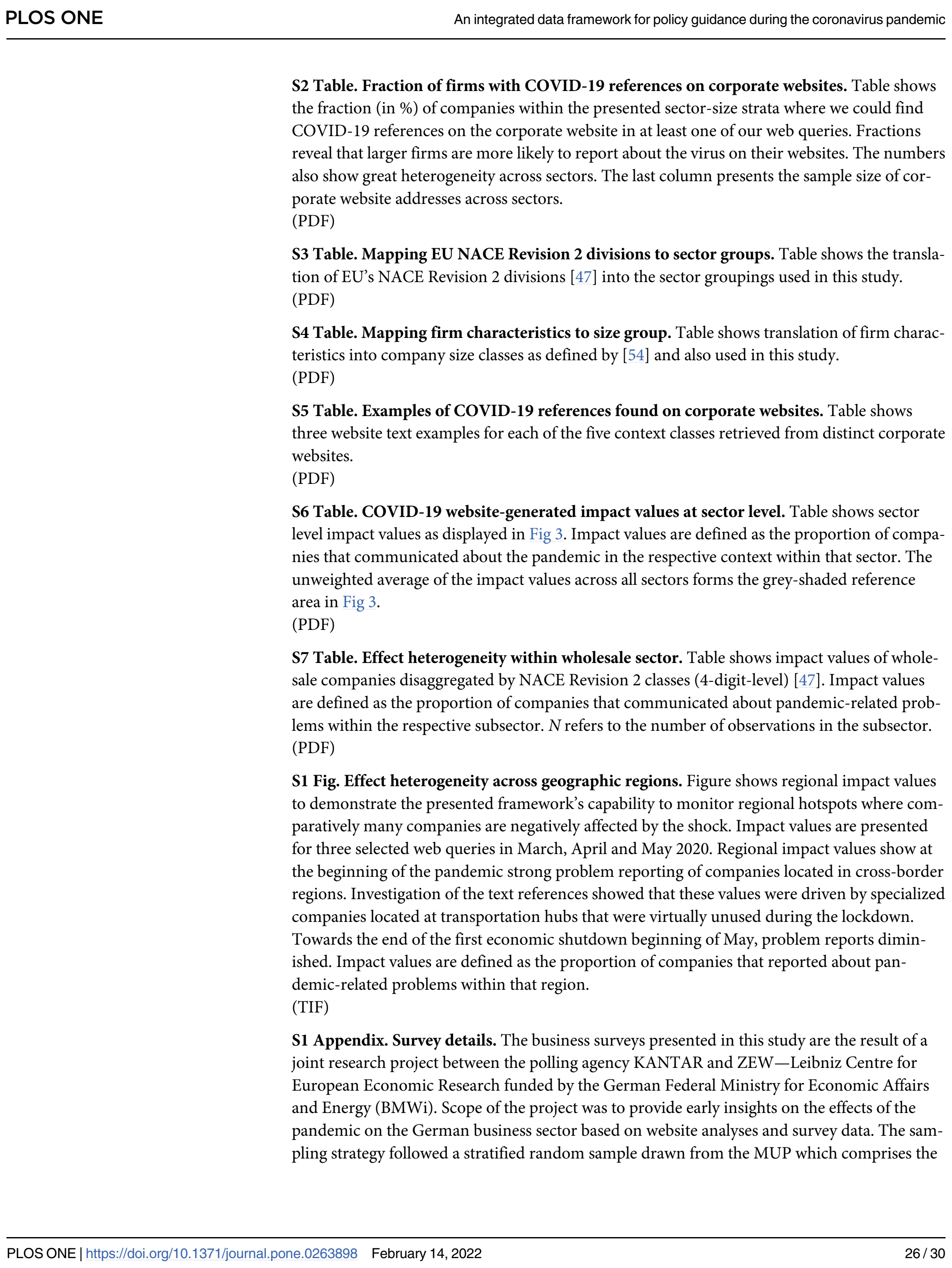
There are, however, limits to our analysis. First and foremost, not all companies have their own corporate website domain, which likely biases our web-based analysis results. Previous studies have shown that URL coverage of German companies is at 46% [32]. Especially among smaller firms the fraction without corporate URL is comparatively high. However, this does not necessarily mean that these companies do not have a corporate online presence at all. Often small and micro firms host corporate profiles on social media platforms to communicate with their stakeholders. It requires further research to detect, access and analyze these online presences to acquire an even more complete picture of corporate communication on the internet in times of economic shocks. Next, company website content is essentially self-reported information that generally bears the risk that firms communicate their current situation overly optimistic (or pessimistic). Interestingly, this study has revealed that in times of economic crises this does not seem to be necessarily the case. On the contrary, we find that close to 70,000 firms reported about problems that they are facing in relation to the pandemic. This equals 35% of all firms that published COVID-19 references on their websites and is substantial given the potential consequences of communicating 'problems' to such a broad audience.

If machine learning-based analysis systems, such as the framework we have presented, indeed find their way into the standard indicator toolkit of policymakers, the question of interpretable (and fair) prediction results will also arise. Complex machine learning models in particular are often deemed as difficult to understand 'black boxes' that do not allow any clear conclusions to be drawn about the factors that are ultimately decisive for predictions and forecasts. In the near future, frameworks like ours will have to integrate aspects of *explainable AI* (see for example Barredo Arrieta et al. [53]) in order to provide decision-makers not only with reliable, but also explainable information as a basis for making informed decisions.

Despite the theoretical drawbacks of our proposed framework, we believe that it is a useful research contribution towards policy guidance that balances timeliness, depth and costs of different data sources. Especially in times of crisis, when sudden shocks cause major disruptions, exploring alternative sources of data is critical to provide timely insights to decision-makers. In this regard, we believe that webdata and other real-time online sources not only serves as a tool to capture business impacts in highly dynamic situations, but also has the potential to support policymakers across a broader spectrum. It is left to future research to explore the value of webdata for policy on a larger scale.

Supporting information

S1 Table. Search terms for querying COVID-19 references on corporate websites. Searches were conducted case insensitive. Spaces in the search terms were treated as wildcards where any two characters instead of the space also led to a match. In this way, we allowed a greater degree of variation in the search for Corona references.
(PDF)



S2 Table. Fraction of firms with COVID-19 references on corporate websites. Table shows the fraction (in %) of companies within the presented sector-size strata where we could find COVID-19 references on the corporate website in at least one of our web queries. Fractions reveal that larger firms are more likely to report about the virus on their websites. The numbers also show great heterogeneity across sectors. The last column presents the sample size of corporate website addresses across sectors.
(PDF)

S3 Table. Mapping EU NACE Revision 2 divisions to sector groups. Table shows the translation of EU's NACE Revision 2 divisions [47] into the sector groupings used in this study.
(PDF)

S4 Table. Mapping firm characteristics to size group. Table shows translation of firm characteristics into company size classes as defined by [54] and also used in this study.
(PDF)

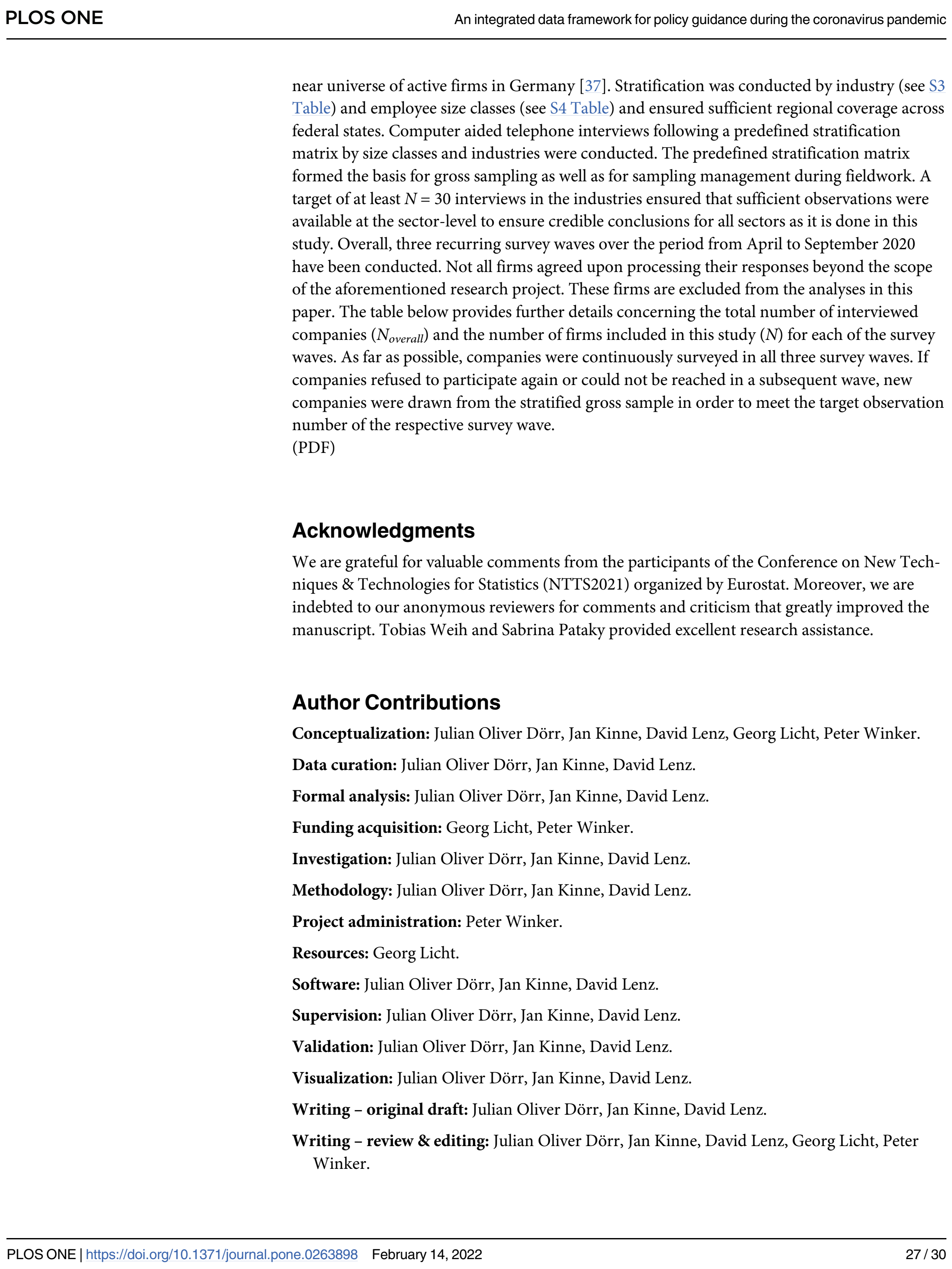
S5 Table. Examples of COVID-19 references found on corporate websites. Table shows three website text examples for each of the five context classes retrieved from distinct corporate websites.
(PDF)

S6 Table. COVID-19 website-generated impact values at sector level. Table shows sector level impact values as displayed in Fig 3. Impact values are defined as the proportion of companies that communicated about the pandemic in the respective context within that sector. The unweighted average of the impact values across all sectors forms the grey-shaded reference area in Fig 3.
(PDF)

S7 Table. Effect heterogeneity within wholesale sector. Table shows impact values of wholesale companies disaggregated by NACE Revision 2 classes (4-digit-level) [47]. Impact values are defined as the proportion of companies that communicated about pandemic-related problems within the respective subsector. *N* refers to the number of observations in the subsector.
(PDF)

S1 Fig. Effect heterogeneity across geographic regions. Figure shows regional impact values to demonstrate the presented framework's capability to monitor regional hotspots where comparatively many companies are negatively affected by the shock. Impact values are presented for three selected web queries in March, April and May 2020. Regional impact values show at the beginning of the pandemic strong problem reporting of companies located in cross-border regions. Investigation of the text references showed that these values were driven by specialized companies located at transportation hubs that were virtually unused during the lockdown. Towards the end of the first economic shutdown beginning of May, problem reports diminished. Impact values are defined as the proportion of companies that reported about pandemic-related problems within that region.
(TIF)

S1 Appendix. Survey details. The business surveys presented in this study are the result of a joint research project between the polling agency KANTAR and ZEW—Leibniz Centre for European Economic Research funded by the German Federal Ministry for Economic Affairs and Energy (BMWi). Scope of the project was to provide early insights on the effects of the pandemic on the German business sector based on website analyses and survey data. The sampling strategy followed a stratified random sample drawn from the MUP which comprises the



near universe of active firms in Germany [37]. Stratification was conducted by industry (see S3 Table) and employee size classes (see S4 Table) and ensured sufficient regional coverage across federal states. Computer aided telephone interviews following a predefined stratification matrix by size classes and industries were conducted. The predefined stratification matrix formed the basis for gross sampling as well as for sampling management during fieldwork. A target of at least $N = 30$ interviews in the industries ensured that sufficient observations were available at the sector-level to ensure credible conclusions for all sectors as it is done in this study. Overall, three recurring survey waves over the period from April to September 2020 have been conducted. Not all firms agreed upon processing their responses beyond the scope of the aforementioned research project. These firms are excluded from the analyses in this paper. The table below provides further details concerning the total number of interviewed companies ($N_{overall}$) and the number of firms included in this study (N) for each of the survey waves. As far as possible, companies were continuously surveyed in all three survey waves. If companies refused to participate again or could not be reached in a subsequent wave, new companies were drawn from the stratified gross sample in order to meet the target observation number of the respective survey wave.
(PDF)

Acknowledgments

We are grateful for valuable comments from the participants of the Conference on New Techniques & Technologies for Statistics (NTTS2021) organized by Eurostat. Moreover, we are indebted to our anonymous reviewers for comments and criticism that greatly improved the manuscript. Tobias Weih and Sabrina Pataky provided excellent research assistance.

Author Contributions

Conceptualization: Julian Oliver Dörr, Jan Kinne, David Lenz, Georg Licht, Peter Winker.

Data curation: Julian Oliver Dörr, Jan Kinne, David Lenz.

Formal analysis: Julian Oliver Dörr, Jan Kinne, David Lenz.

Funding acquisition: Georg Licht, Peter Winker.

Investigation: Julian Oliver Dörr, Jan Kinne, David Lenz.

Methodology: Julian Oliver Dörr, Jan Kinne, David Lenz.

Project administration: Peter Winker.

Resources: Georg Licht.

Software: Julian Oliver Dörr, Jan Kinne, David Lenz.

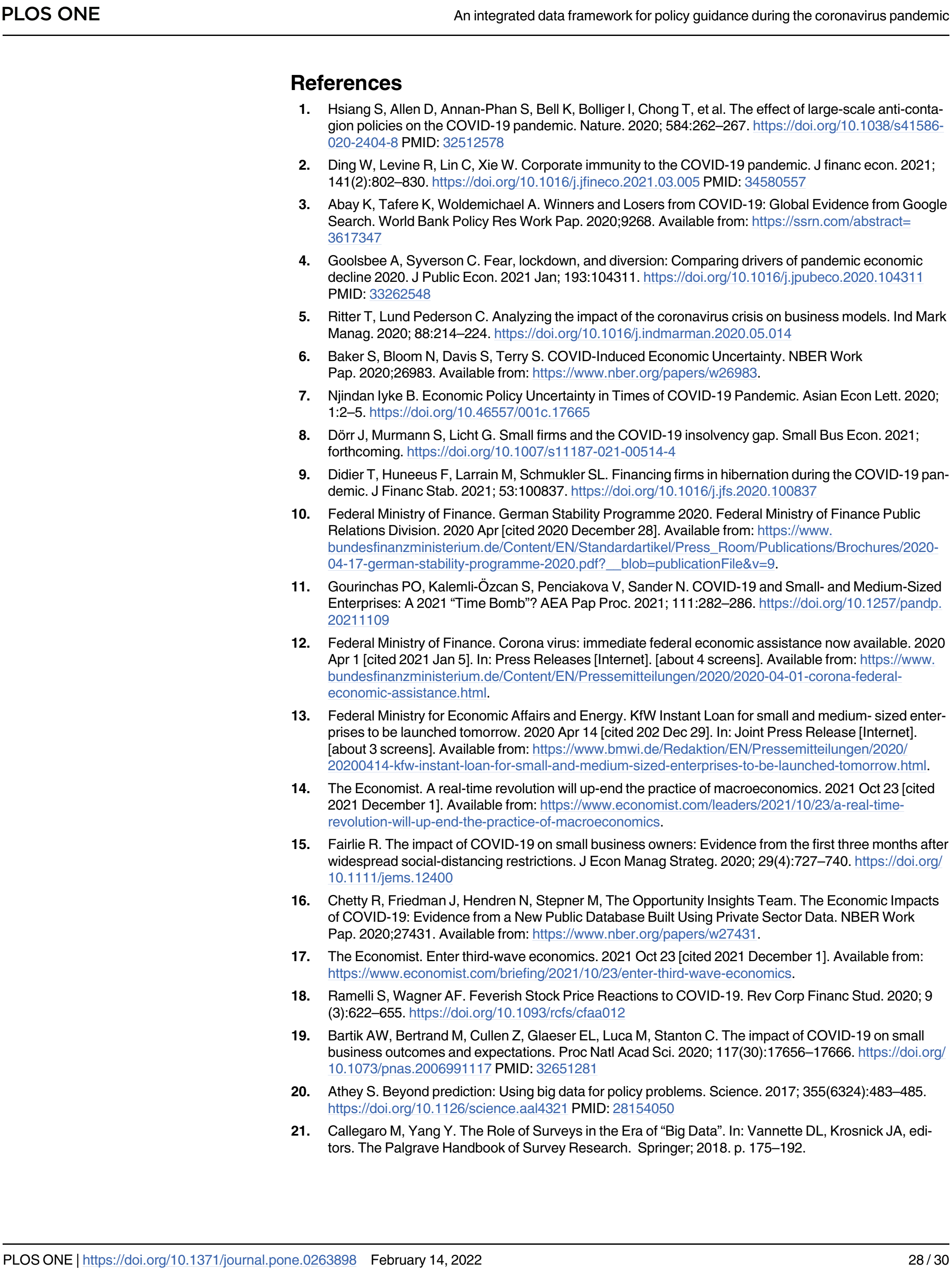
Supervision: Julian Oliver Dörr, Jan Kinne, David Lenz.

Validation: Julian Oliver Dörr, Jan Kinne, David Lenz.

Visualization: Julian Oliver Dörr, Jan Kinne, David Lenz.

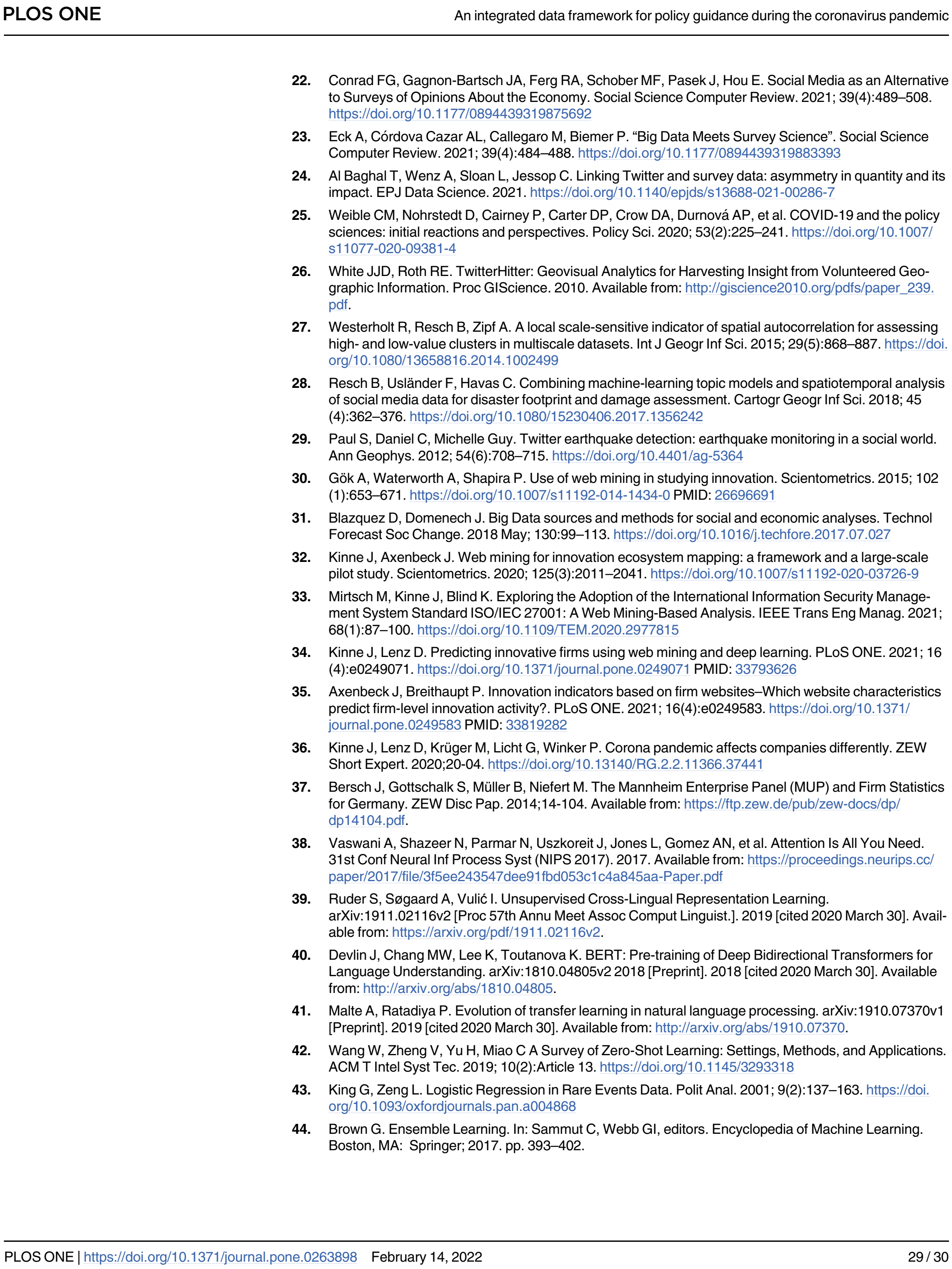
Writing – original draft: Julian Oliver Dörr, Jan Kinne, David Lenz.

Writing – review & editing: Julian Oliver Dörr, Jan Kinne, David Lenz, Georg Licht, Peter Winker.



References

1. Hsiang S, Allen D, Annan-Phan S, Bell K, Bolliger I, Chong T, et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. Nature. 2020; 584:262–267. https://doi.org/10.1038/s41586-020-2404-8 PMID: 32512578
2. Ding W, Levine R, Lin C, Xie W. Corporate immunity to the COVID-19 pandemic. J financ econ. 2021; 141(2):802–830. https://doi.org/10.1016/j.jfineco.2021.03.005 PMID: 34580557
3. Abay K, Tafere K, Woldemichael A. Winners and Losers from COVID-19: Global Evidence from Google Search. World Bank Policy Res Work Pap. 2020;9268. Available from: https://ssrn.com/abstract=3617347
4. Goolsbee A, Syverson C. Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. J Public Econ. 2021 Jan; 193:104311. https://doi.org/10.1016/j.jpubeco.2020.104311 PMID: 33262548
5. Ritter T, Lund Pederson C. Analyzing the impact of the coronavirus crisis on business models. Ind Mark Manag. 2020; 88:214–224. https://doi.org/10.1016/j.indmarman.2020.05.014
6. Baker S, Bloom N, Davis S, Terry S. COVID-Induced Economic Uncertainty. NBER Work Pap. 2020;26983. Available from: https://www.nber.org/papers/w26983.
7. Njindan Iyke B. Economic Policy Uncertainty in Times of COVID-19 Pandemic. Asian Econ Lett. 2020; 1:2–5. https://doi.org/10.46557/001c.17665
8. Dörr J, Murmann S, Licht G. Small firms and the COVID-19 insolvency gap. Small Bus Econ. 2021; forthcoming. https://doi.org/10.1007/s11187-021-00514-4
9. Didier T, Huneeus F, Larrain M, Schmukler SL. Financing firms in hibernation during the COVID-19 pandemic. J Financ Stab. 2021; 53:100837. https://doi.org/10.1016/j.jfs.2020.100837
10. Federal Ministry of Finance. German Stability Programme 2020. Federal Ministry of Finance Public Relations Division. 2020 Apr [cited 2020 December 28]. Available from: https://www.bundesfinanzministerium.de/Content/EN/Standardartikel/Press_Room/Publications/Brochures/2020-04-17-german-stability-programme-2020.pdf?__blob=publicationFile&v=9.
11. Gourinchas PO, Kalemli-Özcan S, Penciakova V, Sander N. COVID-19 and Small- and Medium-Sized Enterprises: A 2021 “Time Bomb”? AEA Pap Proc. 2021; 111:282–286. https://doi.org/10.1257/pandp.20211109
12. Federal Ministry of Finance. Corona virus: immediate federal economic assistance now available. 2020 Apr 1 [cited 2021 Jan 5]. In: Press Releases [Internet]. [about 4 screens]. Available from: https://www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2020/2020-04-01-corona-federal-economic-assistance.html.
13. Federal Ministry for Economic Affairs and Energy. KfW Instant Loan for small and medium- sized enterprises to be launched tomorrow. 2020 Apr 14 [cited 202 Dec 29]. In: Joint Press Release [Internet]. [about 3 screens]. Available from: https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2020/20200414-kfw-instant-loan-for-small-and-medium-sized-enterprises-to-be-launched-tomorrow.html.
14. The Economist. A real-time revolution will up-end the practice of macroeconomics. 2021 Oct 23 [cited 2021 December 1]. Available from: https://www.economist.com/leaders/2021/10/23/a-real-time-revolution-will-up-end-the-practice-of-macroeconomics.
15. Fairlie R. The impact of COVID-19 on small business owners: Evidence from the first three months after widespread social-distancing restrictions. J Econ Manag Strateg. 2020; 29(4):727–740. https://doi.org/10.1111/jems.12400
16. Chetty R, Friedman J, Hendren N, Stepner M, The Opportunity Insights Team. The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data. NBER Work Pap. 2020;27431. Available from: https://www.nber.org/papers/w27431.
17. The Economist. Enter third-wave economics. 2021 Oct 23 [cited 2021 December 1]. Available from: https://www.economist.com/briefing/2021/10/23/enter-third-wave-economics.
18. Ramelli S, Wagner AF. Feverish Stock Price Reactions to COVID-19. Rev Corp Financ Stud. 2020; 9 (3):622–655. https://doi.org/10.1093/rcfs/cfaa012
19. Bartik AW, Bertrand M, Cullen Z, Glaeser EL, Luca M, Stanton C. The impact of COVID-19 on small business outcomes and expectations. Proc Natl Acad Sci. 2020; 117(30):17656–17666. https://doi.org/10.1073/pnas.2006991117 PMID: 32651281
20. Athey S. Beyond prediction: Using big data for policy problems. Science. 2017; 355(6324):483–485. https://doi.org/10.1126/science.aal4321 PMID: 28154050
21. Callegaro M, Yang Y. The Role of Surveys in the Era of “Big Data”. In: Vannette DL, Krosnick JA, editors. The Palgrave Handbook of Survey Research. Springer; 2018. p. 175–192.



22. Conrad FG, Gagnon-Bartsch JA, Ferg RA, Schober MF, Pasek J, Hou E. Social Media as an Alternative to Surveys of Opinions About the Economy. Social Science Computer Review. 2021; 39(4):489–508. https://doi.org/10.1177/0894439319875692

23. Eck A, Córdova Cazar AL, Callegaro M, Biemer P. "Big Data Meets Survey Science". Social Science Computer Review. 2021; 39(4):484–488. https://doi.org/10.1177/0894439319883393

24. Al Baghal T, Wenz A, Sloan L, Jessop C. Linking Twitter and survey data: asymmetry in quantity and its impact. EPJ Data Science. 2021. https://doi.org/10.1140/epjds/s13688-021-00286-7

25. Weible CM, Nohrstedt D, Cairney P, Carter DP, Crow DA, Durnová AP, et al. COVID-19 and the policy sciences: initial reactions and perspectives. Policy Sci. 2020; 53(2):225–241. https://doi.org/10.1007/s11077-020-09381-4

26. White JJD, Roth RE. TwitterHitter: Geovisual Analytics for Harvesting Insight from Volunteered Geographic Information. Proc GIScience. 2010. Available from: http://giscience2010.org/pdfs/paper_239.pdf.

27. Westerholt R, Resch B, Zipf A. A local scale-sensitive indicator of spatial autocorrelation for assessing high- and low-value clusters in multiscale datasets. Int J Geogr Inf Sci. 2015; 29(5):868–887. https://doi.org/10.1080/13658816.2014.1002499

28. Resch B, Usländer F, Havas C. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. Cartogr Geogr Inf Sci. 2018; 45 (4):362–376. https://doi.org/10.1080/15230406.2017.1356242

29. Paul S, Daniel C, Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys. 2012; 54(6):708–715. https://doi.org/10.4401/ag-5364

30. Gök A, Waterworth A, Shapira P. Use of web mining in studying innovation. Scientometrics. 2015; 102 (1):653–671. https://doi.org/10.1007/s11192-014-1434-0 PMID: 26696691

31. Blazquez D, Domenech J. Big Data sources and methods for social and economic analyses. Technol Forecast Soc Change. 2018 May; 130:99–113. https://doi.org/10.1016/j.techfore.2017.07.027

32. Kinne J, Axenbeck J. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics. 2020; 125(3):2011–2041. https://doi.org/10.1007/s11192-020-03726-9

33. Mirtsch M, Kinne J, Blind K. Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis. IEEE Trans Eng Manag. 2021; 68(1):87–100. https://doi.org/10.1109/TEM.2020.2977815

34. Kinne J, Lenz D. Predicting innovative firms using web mining and deep learning. PLoS ONE. 2021; 16 (4):e0249071. https://doi.org/10.1371/journal.pone.0249071 PMID: 33793626

35. Axenbeck J, Breithaupt P. Innovation indicators based on firm websites–Which website characteristics predict firm-level innovation activity?. PLoS ONE. 2021; 16(4):e0249583. https://doi.org/10.1371/journal.pone.0249583 PMID: 33819282

36. Kinne J, Lenz D, Krüger M, Licht G, Winker P. Corona pandemic affects companies differently. ZEW Short Expert. 2020;20-04. https://doi.org/10.13140/RG.2.2.11366.37441

37. Bersch J, Gottschalk S, Müller B, Niefert M. The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. ZEW Disc Pap. 2014;14-104. Available from: https://ftp.zew.de/pub/zew-docs/dp/dp14104.pdf.

38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 31st Conf Neural Inf Process Syst (NIPS 2017). 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

39. Ruder S, Søgaard A, Vulić I. Unsupervised Cross-Lingual Representation Learning. arXiv:1911.02116v2 [Proc 57th Annu Meet Assoc Comput Linguist.]. 2019 [cited 2020 March 30]. Available from: https://arxiv.org/pdf/1911.02116v2.

40. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 2018 [Preprint]. 2018 [cited 2020 March 30]. Available from: http://arxiv.org/abs/1810.04805.

41. Malte A, Ratadiya P. Evolution of transfer learning in natural language processing. arXiv:1910.07370v1 [Preprint]. 2019 [cited 2020 March 30]. Available from: http://arxiv.org/abs/1910.07370.

42. Wang W, Zheng V, Yu H, Miao C A Survey of Zero-Shot Learning: Settings, Methods, and Applications. ACM T Intel Syst Tec. 2019; 10(2):Article 13. https://doi.org/10.1145/3293318

43. King G, Zeng L. Logistic Regression in Rare Events Data. Polit Anal. 2001; 9(2):137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

44. Brown G. Ensemble Learning. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer; 2017. pp. 393–402.

45. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20(1):37–46. https://doi.org/10.1177/001316446002000104

46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. https://doi.org/10.2307/2529310 PMID: 843571

47. European Union. Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regula. Off J Eur Union. 2006; L393:1–39.

48. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance. 1968; 23(4):589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

49. Altman EI. Predicting financial distress of companies: revisiting the Z-Score and ZETA® models. In: Bell AR, Brooks C, Prokopczuk M, editors. Handbook of Research Methods and Applications in Empirical Finance. Cheltenham, UK: Edward Elgar Publishing; 2013. p. 428–456.

50. Creditreform. Creditreform Solvency Index Creditreform Commercial Information. [cited 2020 December 1]. Available from: https://www.creditreform.ro/fileadmin/user_upload/crefo/download_eng/commercial_information/Flyer_Solvency_Index.pdf.

51. Creditreform. Creditreform Commercial Report International Creditreform Commercial Information. 2015 Nov [cited 2020 December 1]. Available from: https://www.creditreform.com/fileadmin/user_upload/CR-International/Bilder/PB_International-Commercial-Report_web.pdf.

52. Chazan G, Fleming S. Germany wields 'bazooka' in fight against coronavirus. Financial Times. 2020 March 13 [cited 2021 May 12]. Available from: https://www.ft.com/content/1b0f0324-6530-11ea-b3f3-fe4680ea68b5.

53. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020; 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

54. European Commission. Commission recommendation concerning the definition of micro, small and medium-sized enterprises. Off J Eur Union. 2003; L124:36–41.

Appendix C

Mapping Technologies to Business Models: An Application to Clean Technologies and Entrepreneurship

Mapping technologies to business models

An application to clean technologies and entrepreneurship

Julian Oliver Dörr*[1,2]

[1]ZEW – Leibniz Centre for European Economic Research, Department of Economics of Innovation and Industrial Dynamics, Mannheim, Germany

[2]Justus Liebig University Giessen, Department of Econometrics and Statistics, Gießen, Germany

May 31, 2022

Abstract

Theory suggests that new market entrants play a special role for the creation of new technological pathways required for the development and diffusion of more sustainable forms of production, consumption, mobility and housing. Unconstrained by past technological investments, entrants can introduce more radical environmental innovations than incumbent firms whose past R&D decisions make them locked into path-dependent trajectories of outdated technologies. Yet, little research exists which provides empirical evidence on new ventures' role in the technological transition towards decarbonization and dematerialization. This is mainly because there is no historical data about entrants' R&D and patenting activities, which are typically used to determine a company's technological footprint. To allow for identification of clean technology-oriented market entrants and to better understand their role as adopters and innovators for sustainable market solutions, this paper introduces a framework that systematically maps new ventures' business models to a set of well-defined clean technologies. For this purpose, the framework leverages textual descriptions of new entrants' business summaries that are typically available upon business registration and allow for a good indication of their technological orientation. Furthermore, the framework uses textual information from patenting activities of established innovators to model semantic representations of technologies. Mapping company and technology descriptions into a common vector space enables the derivation of a fine-granular measure of entrants' technological orientation. Applying the framework to a survey of German start-up firms suggests that clean technology-oriented market entrants act as accelerators of technical change: both by virtue of their existing products and services and through a high propensity to introduce additional environmental innovations.

Keywords: Clean technologies, technological orientation, environmental innovation, sustainable entrepreneurship, text modeling, natural language processing

JEL: C38, O13, Q55

*E-mail: julian-doerr@gmx.de

The author is grateful for valuable comments from Peter Winker, Hannes Marks and Tobias Weih which improved an earlier version of the manuscript.

1 Introduction

Given anthropogenic climate change and the rapid depletion of the remaining carbon budget that limits global warming to a manageable level, the development and diffusion of clean, environmentally sound technologies play an increasingly important role in accelerating the transition to a low-carbon economy. This has been acknowledged in the Paris Agreement of 2015 which stresses the 'importance of [. . .] technology development and transfer in order to improve resilience to climate change and to reduce greenhouse gas emissions' (United Nations 2015, p. 14). According to the United Nations (2015), this technological shift requires innovations and increased investment in more sustainable forms of production, consumption, mobility and housing. This clearly brings entrepreneurs as a crucial source of innovation to the fore. Sustainable entrepreneurship, in particular, has become an important stream of research to understand the role of dedicated business models for the technological transition to decarbonization and dematerialization.

While research on sustainable entrepreneurship largely agrees that environmental innovations are inherent to both established companies and new market entrants (Hockerts & Wüstenhagen 2010; Schaltegger & Wagner 2011; Gast et al. 2017), there is relatively little empirical work that specifically analyzes the transitional impact of the latter group. Yet, from a theoretical standpoint, new ventures are attributed a special role for the creation of new technological pathways. Unconstrained by previous investment decisions, entrants can introduce more radical environmental innovations than incumbent firms. In this way, theory suggests that entrants act as accelerators for the diffusion of clean technologies (Hockerts & Wüstenhagen 2010; Fichter & Clausen 2013) and may also help to overcome transition inertia among incumbents (Diekhof 2015).

Empirically, firm-level indicators that reflect a company's technological footprint are necessary to identify which role different types of companies - e.g. established firms in contrast to entrants - play in the diffusion of new technologies. Typically, technology and innovation research relies on patent and R&D information to determine a firm's technological profile (Archibugi & Pianta 1996; Aharonson & Schilling 2016).[1] However, unlike established companies, there exists no historical track record on R&D investments for new business ventures, and patent activities are also rare among start-ups (Graham & Sichelman 2008; Helmers & Rogers 2011). The lack of such innovation-related data makes it inherently difficult to empirically narrow down market entrants' technology usage and innovation capability. Moreover, existing classification statistics such as industry affiliation, tend to be too broad to capture a subtle construct such as a firm's ori-

[1] Of course, there are also innovation surveys which, apart from common survey problems such as cost intensity and non-response, appear impractical for measuring company-specific technology portfolios from a very broad spectrum of different technologies. Nonetheless, see Comin et al. (2020) for a recent attempt to survey companies across 287 distinct technologies.

entation towards environmentally-sound technologies. For these reasons, research suggests that understanding the impact of new ventures on accelerating sustainable market transformations is much more a question of 'predictive, modeling-based, ex-ante evaluation than of retrospective, experienced-based, ex-post evaluation which applies to established companies' (Trautwein 2021, p. 3). In other words, for companies that are new to the market, only information available at or shortly after the company's foundation can be used to predict its transformational capability with respect to the development and diffusion of clean technology solutions.

This paper follows this predictive approach by focusing on new ventures' orientation towards clean technologies as ex-ante indicator of their contribution to the transition towards more sustainable market standards. For this purpose, the paper leverages observable and detailed business summaries that new ventures are typically obliged to report upon business registration.[2] The legal obligation to publish a business purpose provides researchers and policymakers not only with fine-grained information about companies' original business activities but also gives a good indication whether specific types of technologies are relevant to their business model. This is demonstrated by the following example of a business summary of a firm from the geothermal energy sector.

> 'Manufacture, sale, maintenance and repair of heat pumps and other technical equipment, in particular for generating thermal energy.'[3]

Based on this textual source of firm-level information, this study shows that it is possible to construct an indicator that reflects a new venture's potential to contribute to the diffusion of a specific technology by mere virtue of its technological orientation. For this purpose, the paper leverages recent advances in the field of natural language modeling to create a mapping of a technological system and to use market entrants' business descriptions to determine their position within this system. In this way, it becomes possible to measure how closely a firm's business model is oriented towards a particular technology: a measure referred to as technological proximity in the remaining of the paper.

The scope of this study is twofold. First, to the best of my knowledge, the proposed measure of technological proximity is the first one which maps business models to a fine-grained level of distinct technologies. Most importantly, the indicator is applicable to market entrants which typically lack track records of alternative technology and innovation indicators. While in theory the approach is highly flexible and allows to position *any* kind of company within *any* kind of technology system, this study applies the approach to position market entrants within a system

[2] In Germany, for example, limited liability companies are legally obliged to state their business purpose as part of the business registration process. See Limited Liability Companies Act (Section 3 (1) No. 2 GmbHG) and Stock Corporation Act (Section 23 (3) No. 2 AktG) for the legal basis of the obligation.

[3] Business description retrieved from the Mannheim Enterprise Panel (MUP) which contains various firm characteristics for the near universe of German companies including textual information on the firms' business purpose as retrieved from the German company register (Bersch et al. 2014).

of well-defined clean technology areas. More specifically, as second contribution of this paper, the framework is applied to a representative survey of German start-up firms in order to investigate the environmental innovation capability of clean technology-oriented market entrants as well as the environmental impact of their products and services. Empirical results suggest that clean technology-oriented firms' products and services have positive environmental effects for their customers in terms of emission reduction, energy efficiency and higher levels of recyclability. Moreover, a higher cleantech orientation at founding predicts a higher propensity to introduce environmental innovations over the course of the venture's lifetime. This suggests that cleantech ventures have a special role to play in the technological transition towards decarbonization and dematerialization: besides their existing products and services building on clean technology solutions, they are also drivers of innovation by introducing new products and services that have a superior environmental footprint and fundamentally differ from their existing product portfolio. These results are in line with theory on new technological path creation triggered by market entrants.

The remaining of the paper is structured as follows. Section 2 discusses the role of new ventures in the technological transition towards sustainable market transformations from a theoretical perspective. In doing so, it relates the study to existing literature on technological path dependency as well as to the theory on externalities in the diffusion of sustainable technologies and environmental innovations. Section 3 introduces the methodological framework used to develop a fine-grained measure of technological orientation at the firm-level. To demonstrate the usefulness of the proposed framework, Section 4 uses the novel measure to assess the clean technology orientation for a representative sample of German start-up firms and analyzes how clean technology-oriented business models relate to the firm's environmental performance. Section 5 concludes.

2 Theoretical background

A key driver of technological change and transformation is the innovative capacity of entrepreneurship (Audretsch et al. 2002; Acs & Audretsch 2005). The technological transition towards decarbonization and dematerialization requires entrepreneurial solutions with a dedicated technological orientation. In literature, sustainable entrepreneurship is seen as an important accelerator of sustainability oriented innovations and technological advances required to leverage cleaner and more sustainable standards of production, transportation and energy generation (Cohen & Winn 2007; Kant 2018; Leendertse et al. 2021). Research largely agrees that sustainable entrepreneurship is inherent to very different forms of organizations. Most notably, it is not exclusive to small innovative entrants, but it is also assumed by large established incumbents (Hockerts & Wüstenhagen 2010; Schaltegger & Wagner 2011; Gast et al. 2017) with much of its transformative power depending on the interaction dynamics between the two (Schaltegger et al. 2016). However, from a theoretical standpoint, there are important differences between established and start-up firms when it comes to their role as cleantech accelerators.

Most notably, incumbent firms are constrained by their past technological investments and the current technology regime in which they operate (Patel & Pavitt 1997; Aghion et al. 2016). Stuck in technological path dependencies, this makes them often inclined to preserve their rents associated with their existing technology portfolio which often builds on inferior and outdated sustainability standards (Unruh 2000; Bohnsack et al. 2014). When facing technological discontinuities, their willingness to implement disruptive innovations is generally limited. Rather, they focus on incremental technological advancements of their existing technology stock (Henderson 1993; Unruh 2000; Smink et al. 2015; Schaltegger & Wagner 2011). In the context of transitioning to a low-carbon economy, incumbents' path dependency, thus, tends to promote a 'locked-in' state of carbon-intensive technological standards and a reluctance to drastically switch to low-carbon technologies (Benner 2009; Dijk et al. 2016; Sick et al. 2016). So even if established firms engage in environmental innovation activities, their incremental nature does not target at accelerating sustainability transformation but rather at preservation of market power.

New entrants, on the contrary, are not constrained by previous investment decisions and are thus free from innovation rigidity due to technological path dependencies. This allows them to tackle market opportunities in a more creative and disruptive manner (Unruh 2000; Schaltegger & Wagner 2011), especially in energy-intensive industries where technological lock-in tends to be particularly strong (Erickson et al. 2015). Therefore, many scholars see a key role in new ventures to spark environmental innovations in order to accelerate the development and diffusion of clean technologies (Cohen & Winn 2007; Fichter & Clausen 2013; Horne & Fichter 2022). Most notably, their search for sustainable market solutions, which often begins in niche markets, has the potential to trigger clean innovation activities among otherwise rigid incumbents in mass markets (Hockerts & Wüstenhagen 2010; Diekhof 2015). It is this multiplier potential which explains market entrants special role as accelerators of clean innovations.

However, environmental innovations generally suffer the widely studied double externality problem (Rennings 2000) which affects incumbents and entrants alike. On the one hand, sustainable entrepreneurs face the risk of not being able to fully internalize the value of their technological developments in light of knowledge spillovers to competitors. On the other hand, clean innovation efforts are also hampered by the lack of full internalization of the environmental costs caused by companies whose business models are based on carbon-intensive processes and ecologically inferior technologies. This double burden presents barriers for innovative entrepreneurs to enter clean technology markets in the first place and calls policy to de-risk and incentivize their decisions to both enter the market and to innovate (Malen & Marcus 2017; Goldstein et al. 2020). Consistent with literature on directed technological change, which has shown that policy can successfully promote clean innovation activities among incumbent firms (Acemoglu et al. 2012; Aghion et al. 2016; Calel & Dechezleprêtre 2016; Hötte 2020), I argue that policy instruments that specifically target the creation of new cleantech firms have great potential to further accelerate the diffusion of clean technologies. In fact, the few empirical research papers on new entrants in cleantech suggest that policymakers play an important role in fostering cleantech start-ups. Covering 24 OECD countries, Cojoianu et al. (2020), for example, show that more stringent environmental policy regimes make it easier for newly founded cleantech ventures to attract investments. This facilitates their establishment in the market and may favor higher technology standards in terms of sustainability in the long term. Moreover, for the U.S., Doblinger et al. (2019) show that technology development alliances between government organizations such as national laboratories and cleantech start-ups increase the innovation activities of the latter.[4]

To effectively direct technical change into desirable pathways, policymakers need to understand to what extent new ventures engage in the adoption and advancement of specific clean market solutions and which cleantech areas are barely tackled by entrepreneurs. In other words, it requires a framework that allows for a mapping of clean technologies and entrepreneurial activities to disclose the interplay between technological advancement and entrepreneurship. The scope of this study is to provide such a mapping framework which allows to tackle several policy needs required to direct and monitor technological change towards sustainable market transformations. In this context, the framework serves as useful tool for policymakers to scan, for example, business registries for clean technology-oriented entrepreneurs. This can be an effective way to direct R&D subsidies, tax incentives and other start-up support towards ventures with high potential to accelerate technical change by mere virtue of their business models' technological orientation. Most notably, with the proposed framework, this selection procedure is possible early on in the lifetime of potential candidates, i.e. upon their business registration.

The paper shows that the framework successfully identifies market entrants with strong environmental performance. This not only underpins the framework's usefulness as a policy tool. It also suggests that clean technology-oriented entrants act as accelerators in the technological transition towards decarbonization and dematerialization: both, by virtue of their existing

[4] Note that Doblinger et al. (2019) obtain information about cleantech start-ups from the i3 Cleantech Group database (Cleantech Group 2022) which comprises information on cleantech firms collected by a team of industry and technology experts. Cojoianu et al. (2020) identify cleantech ventures by manually examining the websites of those start-ups which have been tagged with a green energy label in the proprietary Crunchbase dataset. Both approaches require labor-intensive manual selection processes that are prone to subjective bias and lack a codified approach to identifying clean technology-oriented entrants.

products and services and by a high propensity to introduce additional environmental innovations. The following section presents the technology mapping framework in detail. In addition to the methodological details on which the framework is built, it also introduces distinct domains of clean technology solutions that form the starting point for creating a mapping of a clean technology system.

3 Measuring Technological Orientation

Technological change and entrepreneurship are two interdependent concepts. Following (Audretsch et al. 2002, p. 157), 'what defines the entrepreneur is the ability to move technology forward into innovation'. A new technology will only diffuse if it has economic value, i.e., if it is put into productive use by someone. The economic application of a new technology by entrepreneurs is thus a necessary condition for the diffusion of the technology and, at the aggregate level, for technological change. This motivates to measure technology usage at firm-level to capture both direction and drivers of technological change. In light of directed technical change, this serves as useful policy tool. It effectively allows to identify entrepreneurial ventures whose technological orientation favors the desired technological pathway. Focusing on the technological transition towards higher levels of decarbonization and dematerialization, this section starts with the definition of a well-defined set of clean technology fields followed by a detailed discussion how a fine-grained measure of technological orientation at firm-level can be derived by means of textual innovation data.

3.1 Mapping of clean technology system

In this paper, 'clean technologies' refer to any process, product or service that aims at reducing negative environmental impacts. This comprises environmental protection and climate change mitigation measures, the sustainable use of natural resources and the use of goods that are modified to be less material- and energy-intensive than the industry standard (dematerialization). Another field of clean technologies is the reduction of anthropogenic emissions and pollution (decarbonization). This includes a wide range of different technologies, from renewable energy generation to carbon capture technologies to clean water technologies, all of which find application across different sectors and create different markets for companies to operate in. To define clearly distinguishable areas of clean technologies, this paper follows the European Patent Office (EPO) classification scheme for green technologies, which 'cover[s] all significant climate change mitigation technologies [. . .] in energy, carbon capture, transport, buildings, waste, energy-intensive industries and smart grids' (United Nations Environment Program & European Patent Office 2015, p. 8). Furthermore, cleantech categories employed in previous literature (Doblinger et al. 2019; Cojoianu et al. 2020) and those published by the Cleantech Group[5], a leading research and consulting agency in the market for clean technologies, are also incorporated. The final list consists of 10 different areas of clean technologies and can be found in Table 1 along with a specific technology example for each area.

Figure 1: Clean technology fields

	Clean technology field	**Technology example**	**Corresponding CPC classes by EPO**
1	Technologies for the adaption to climate change (Adaption)	Genetically modified plants resistant to drought	Y02A 10, Y02A 30-60, Y02A 90, Y02B 80
2	Battery storage and fuel cells (Battery)	Fuel cell technologies in production processes	Y02B 90/10, Y02E 60/30, Y02E 60/32, Y02E 60/34, Y02E 60/36, Y02E 60/50, Y02E 60/30, Y02P 90/40, Y02P 90/45, Y02P 90/50, Y02T 90/40
3	Biofuel technologies (Biofuels)	Algae biomass	Y02E 50, Y02T 10/30
4	Carbon capture, storage and sequestration (CCS)	Enhanced coal bed methane recovery	Y02C 10, Y02C 20, Y02P 40/18, Y02P 70/10, Y02P 90/70
5	Energy efficiency (E-efficiency)	Insulation technologies inhibiting radiant heat transfer	Y02B 20-50, Y02B 70, Y02B 90 (Y02B 90/10), Y02D 10, Y02D 30, Y02D 70, Y02E 20, Y02E 40, Y02P 80
6	Renewable energy generation (Generation)	Generation of geothermal energy	Y02E 10, Y02E 30, Y02B 10, Y02P 10/20, Y02P 20/143, Y02P 20/582, Y02P 20/584, Y02P 70 (except Y02P 70/10)
7	Grid and power conversion (Grid)	Smart grids	Y02E 60/10, Y02E 60/13, Y02E 60/14, Y02E 60/16, Y02E 70, Y02T 10/70, Y04
8	Low carbon materials and manufacturing (Materials)	Technologies to replace cement by fly ash in concrete production	Y02P 10-40 (except Y02P 10/20, Y02P 20/143, Y02P 20/582, Y02P 20/584), Y02W 90
9	Electric vehicles and low carbon mobility solutions (Mobility)	Ultracapacitors for efficient electric vehicle charging	Y02T 10 (except Y02T 10/30, Y02T 10/70), Y02T 30, Y02T 50, Y02T 70, Y02T 90 (except Y02T 90/40)
10	Water and wastewater treatment (Water)	Technologies for the production of fertilisers from the organic fraction of waste or refuse	Y02A 20, Y02W 10, Y02W 30

Note: Clean technology fields form the basis for deriving a mapping between specific clean technologies and business models. Patent documents labeled with the corresponding Cooperative Patent Classification (CPC) classes by the European Patent Office (EPO) as listed in the last column are used to derive semantic representations of the respective clean technology field.

These technology fields form the basis for mapping a clean technology system. The mapping approach makes use of semantic information about the underlying technologies as retrieved from a large corpus of technical patent texts. In essence, the semantic mapping consists of two steps:

(i) Modeling of semantic technology *descriptions* for each of the above clean technology fields. A semantic technology description is best described as a sequence of technological terms which refer with high probability to the focal technology. These word-based technology descriptions are derived empirically from a large corpus of expert-labeled patent abstracts.

(ii) Leveraging the semantic technology description to a *vector representation* by means of text embedding models. This step shifts the word-based technology description to a context-based numerical vector which determines the technologies' position within technological

[5] https://www.cleantech.com

space.

In the following, these two steps and the underlying methods will be introduced in more detail.

From patents to semantic technology descriptions

This study uses an expert-labeled corpus of patent abstracts as the basis for constructing semantic representations for the different clean technology fields. Overall, the corpus comprises more than 550,000 patent documents filed by patent applicants located in Germany.[6] Given the technical content of patent documents, semantic patent analysis poses a natural starting point for technology-related research such as technology forecasting (Guo et al. 2016; Zhang et al. 2016; Song et al. 2017; Chen et al. 2017; Hwang & Shin 2019), technology roadmapping (Lee et al. 2008; Choi et al. 2013; Geum et al. 2015; Zhang et al. 2016) and more recently for analyzing technology profiles (Suominen et al. 2017) and business method innovations within firms (Moehrle et al. 2018).[7] This study leverages the textual content of patents to derive semantic descriptions of technologies, i.e. to model technologies semantically. Besides the textual content of the patent documents, the paper also makes use of patents' metadata which are typically assigned as part of the patent's granting process. Most importantly, it uses the patent's Cooperative Patent Classification (CPC) classes which helps patent examiners to group inventions by technical area. According to the EPO, at its finest level of granularity, there are about 250,000 distinct CPC labels that map patents to underlying technologies (European Patent Office 2020). Most importantly, for the case of clean technologies, the CPC scheme incorporates a class for climate change mitigation technologies, the so called Y02 taxonomy, which allows for the identification and classification of patents whose invention relate to the clean technology fields introduced above.[8]

Acknowledging that clean technologies span various technical fields relevant in very different industrial sectors, the Y02 taxonomy has been introduced as a complementary scheme to the already existing classification schemes at EPO.[9] For this reason, cleantech patents are typically not only assigned to one CPC label that uniquely relates to a single technology field. Instead,

[6] German patent filers are selected because the assessment of new ventures' proximity to the different cleantech fields in Section 4 focuses on German start-ups. As country of the *Energiewende*, Germany has long been regarded as a regulatory pioneer with regard to its commitment to a low-carbon economy and its promotion of eco-innovative technologies. With this form of directed technical change, it is expected that policy has also incentivized the creation of new ventures in the clean technology domain. Thus, it is seen as likely that a representative sample of German start-ups will contain cleantech ventures.

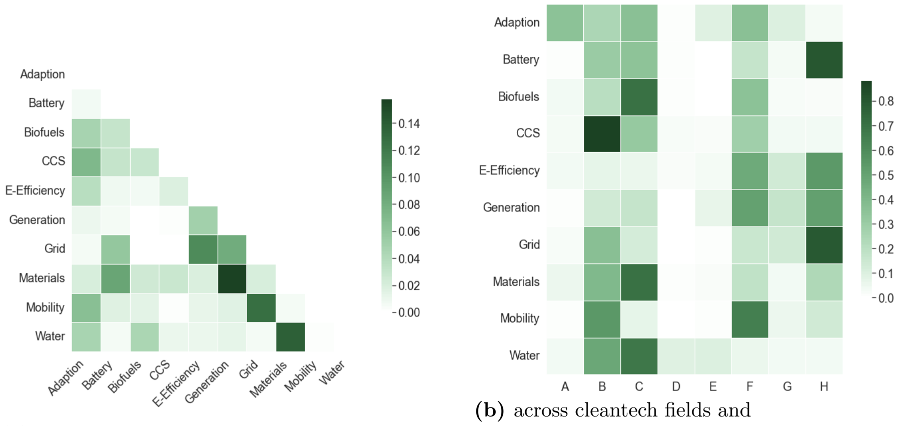
[7] Note that these studies are limited to companies that file patents, which is rarely the case for market entrants.

[8] At its least granular level, the Y02 taxonomy spans eight different subclasses. The definition of the clean technology fields derived in this paper closely follows these subclasses. The exact mapping between cleantech fields used in this study and Y02 labels by EPO can be found in Table 1.

[9] In fact, the Y02 class is the result of an unprecedented effort by EPO to assess all patents ever filed with EPO that are related to clean technologies. Both specialized patent examiners from EPO together with outside experts from the various clean technology fields jointly developed the Y02 taxonomy in order to ensure its validity. Today, more than 3.2M patent documents fall under the Y02 scheme which is why it is seen as the most accurate labeling of clean technology patents available and the international standard for clean innovation studies (Calel & Dechezleprêtre 2016).

most patent documents are co-labeled with CPC classes which refer to different cleantech fields and non-cleantech related technology fields. This becomes apparent in Figure 2 which shows that most patents have some degree of technical complementary and are thus applicable to different technology fields. This makes it challenging to retrieve those technical terms which closely resemble the technology field of interest. In order to derive technology descriptions from the technical terms of the patent texts, a statistical procedure is required to disambiguate which words refer to which technology with highest probability.

Figure 2: Complementarity of cleantech fields in patent corpus



(a) across different cleantech fields

(b) across cleantech fields and non-cleantech-related CPC classes

Note: Complementarity indicates the percentage of patents assigned to the cleantech field on the horizontal axis that are also assigned to (a) the cleantech field on the vertical axis as well as to (b) the non-cleantech-related CPC classes A-H.

Statistically, this translates into the goal to model a probability vector, δ_t, over the corpus' vocabulary, V, for each of the technology fields, t.[10] The intuition here is that technological terms accompanying patents that are relatively frequently assigned to a particular technology field semantically circumscribe that technology. Due to the co-labelling of patent documents, none of these technical terms is exclusive to a technology field, but modeling technology-specific probability vectors over all terms allows to disentangle the terms' relative importance of circumscribing a particular technology field. In other words, the word probability vectors δ_t for all t 'distribute' the corpus' technical terms to technologies. A common approach to derive δ_t is provided by probabilistic topic modeling such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003). LDA assumes that the patent corpus arose from a generative process that is defined by a joint probability distribution over both the observed terms in each patent document but also hidden variables such as the probability vector over the vocabulary for each technology (Blei 2012). As a completely unsupervised algorithm, LDA does not allow patent labels to be incor-

[10] In other words, a technology description is defined as probability distribution over the fixed vocabulary of the patent corpus.

porated into the algorithm. Therefore, this paper follows Ramage et al. (2009)'s Labeled Latent Dirichlet Allocation (L-LDA) extension which adds supervision to the algorithm by restricting the generative process to only consider technology fields which accompany the patents through their CPC labels. So, with the patent corpus, D, that consists of P distinct patent abstracts each of length N_p the generative process can be modeled as follows.

1. For each technology field $t \in \{1, \ldots, T\}$: generate the word distribution from a Dirichlet prior $\delta_t \sim Dir(\beta)$
2. For each patent $p \in \{1, \ldots, P\}$: generate a patent-specific technology distribution from another Dirichlet prior $\lambda_p \sim Dir(\alpha_p)$. This is where the algorithm includes supervision since parameter α_p restricts the Dirichlet to only consider the technology fields which accompany the patent through their CPC labels.[11]
3. For each of the word positions p, n, with $p \in \{1, \ldots, P\}$ and $n \in \{1, \ldots, N_p\}$:
   (a) generate the technology assignment according to $z_{p,n} \sim Multinomial(\lambda_p)$[12]
   (b) and select words according to $w_{p,n} \sim Multinomial(\delta_{z_{p,n}})$

This way the generative process fully specifies both the observed words from the patent abstracts, w, *and* hidden random variables that cannot directly observed from the corpus (Blei 2012). These hidden variables comprise the distribution of technology fields over patent abstracts, λ_p,[13] the technology assignment for the nth word in patent p, $z_{p,n}$, and, most importantly in the context of this study, the word distribution for each clean technology field, i.e. δ_t. The above specification of the generative process corresponds to the joint probability distribution

$$p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P}) = \prod_{t=1}^{T} p(\delta_t) \prod_{p=1}^{P} p(\lambda_p) \left(\prod_{n=1}^{N_p} p(z_{p,n}|\lambda_p) p(w_{p,n}|\delta_{1:T}, z_{p,n}) \right). \tag{1}$$

The statistical learning problem to obtain technology-specific word distributions from the observed patent abstracts is to infer the posterior distributions $p(\delta_t)$, i.e., to derive the marginal distribution $p(\delta_t)$ from the above joint probability distribution. Following Ramage et al. (2009), this study uses Gibbs sampling to derive the posterior word distributions.

A semantic technology description, X_t, is then defined by the technical terms from the patent corpus whose probability of referring to technology t is highest. For example, the word probability distribution for Carbon Capture and Storage (CCS) technologies, $p(\delta_{CCS})$, yields

[11] In LDA, all patent documents would share the same set of technologies, but each patent exhibits those technologies with different proportion. Unlike LDA, the generative process used in this study (L-LDA) restricts the model to only consider the technology fields which accompany the patent through their CPC labels. It does so by modeling the technology field attribution, determining α_p, via a simple Bernoulli prior for each of the T technology classes (see Ramage et al. (2009) for details).

[12] Similar to α_p, the generation of $z_{p,n}$ is restricted to technology fields that accompany the patents.

[13] While the technology fields relevant to a patent are observable through its CPC labels, the patent's *proportion* attributable to each of the fields is hidden.

the following semantic technology description

$$\underset{(1\times Q)}{X_{CCS}} = \langle \text{gas, absorption, dioxide, carbon}, \ldots, \text{scrub, seperation, desorption} \ldots \rangle$$

with the terms ordered by descending probability.[14]

As sequence of technical terms, the semantic technology descriptions convey an intuitive understanding of the technology they are intended to describe. For example, the word ‘gas’ by itself gives little indication of CCS technologies. But ‘gas’ taken together with terms like ‘absorption’, ‘carbon’, and ‘scrub’ provide a high context that can closely be inferred to CCS technologies.[15]

From semantic technology descriptions to technology embeddings

Text embedding models are a common method for converting word sequences into a vector format while preserving the context of the sequence. Text embedding models build on the concept of word embeddings which are dense vector representations of words that allow words with similar meaning to have a similar representation in vector space. The core idea in deriving word embeddings is to exploit information about the co-occurrence of words, i.e. the appearance of two words in close proximity in large text corpora. In recent years, this has been a very active research field, which has led to major advances in network architectures (see Wang et al. (2020) for an overview) to derive highly contextualized word and text embeddings. This paper makes use of a pretrained text embedding model that is based on the seminal Bidirectional Encoder Representations from Transformers (BERT) network architecture (Devlin et al. 2018).[16] Specifically, I use a pretrained version of Sentence-BERT (SBERT) (Reimers & Gurevych 2019) to encode the semantic technology descriptions as fixed-size, dense vectors which I refer to as technology embeddings in the remaining of the paper.

$$\begin{tikzcd}
\underset{(1\times Q)}{X_{CCS}} = \langle \text{gas, absorption, dioxide, carbon}, \ldots, \text{scrub, seperation, desorption} \ldots \rangle \arrow[d, "\text{SBERT}"'] \\
\underset{(1\times 384\,\forall Q)}{X_{CCS}} = [0.479, -0.016, \ldots 0.483, -0.347]'
\end{tikzcd}$$

Note that the last layer in a SBERT network is a pooling operation that averages all word embeddings and thus produces fixed-size output vectors regardless of the length of the input

[14] Note that the final number of technical terms used to model the semantic technology descriptions, Q, is treated as hyperparameter whose optimal value is determined empirically (see Section 3.3).

[15] See Table 7 in the Appendix for the most relevant technical terms for all clean technology fields.

[16] Unlike previous language models, BERT's network architecture and training objective allows it to derive word embeddings based on the context given before (on the left side of) the focal word *and* after (on the right side of) the focal word (Wang et al. 2020). Thus, BERT no more treats word sequences as unidirectional left-to-right sequence but as *bidirectional* sequence of word dependencies.

sequence (Reimers & Gurevych 2019). In the specification of this study, the fixed size vector has length 384.

Conducting the encoding for all of the 10 clean technology descriptions yields a mapping of the clean technologies in semantic vector space. In the next section, I show how to place companies into the same vector space based on their business descriptions. I then propose a distance measure between technology and company vectors to determine how 'close' or 'distant' a firm is positioned to each of the technologies. Moreover, Section 3.3 shows that the discriminative 'quality' of the measure depends on the number of words, Q, that are used to model the semantic technology description. Ultimately, this number is determined empirically using a technology-labeled dataset of business descriptions.

3.2 Deriving a technological proximity measure

In order to position companies within the clean technology system, I use the same pretrained SBERT model to derive vector representations of each firms business summary. In this way, it becomes possible to position companies within the system of clean technologies and ultimately to determine their proximity (distance) to each of the technologies. Sentence-BERT (SBERT) has been fine-tuned on semantic textual similarity data, i.e., pairs of word sequences that have been labeled as 'contradiction', 'entailment' or 'neutral' (Reimers & Gurevych 2019). This makes SBERT highly suitable for the derivation of a technological proximity measure where the goal is to determine whether a new venture's business description is 'close' to a certain technology description or rather 'distant' from it. If two word sequences (texts) consist of distinct words but share a similar context, SBERT will encode the sequences into similar vector representations. For example, a description of a new venture, Y_i, that has specialized in CCS technologies may look as follows:

> 'Development and licensing of direct air capture technology that safely and permanently removes CO2 from the air.'

Although there is no direct word overlap between X_{CCS} and Y_i, the word embeddings of some of the words in both descriptions are likely to be highly correlated. For instance, 'gas', 'carbon', 'dioxide' in X_{CCS} and 'air' and 'co2' in Y_i are likely to be close to each other in vector space as in very large corpora these words tend occur in close proximity to each other relatively often. The same applies to 'absorption', 'desorption' in X_{CCS} and 'capture' and 'remove' in Y_i.

This paper proposes cosine similarity to quantify a ventures technological orientation towards a specific technology.

$$\textsc{TechProx}_{t,i} := sim(X_t, Y_i) = cos(\theta_{t,i}) = max\left(0, \frac{\bar{X}_t \bar{Y}_i}{|| \bar{X}_t |||| \bar{Y}_i ||}\right) \in [0,1] \qquad (2)$$

Cosine similarity as measure of semantic similarity between two texts is well documented (see for example Chandrasekaran and Mago (2021) for a recent survey). Intuitively, if the angle between a company and a technology embedding is small, both vectors point into similar directions in the clean technology system which means that they share similar context (i.e., they share contextually similar words). The more a company's business model relates to a specific clean technology, the higher the semantic overlap between company and technology description and, thus, the closer TechProx moves towards its maximum value of 1. If the words in the company description are however contextually independent from the technology description's words, TechProx takes on a value close to 0 indicating that the firm's business model is not related to the respective clean technology.[17]

3.3 Validating the technological proximity measure

Up to this point, this section has shown how patent texts can be used to map a system of different technologies in vector space. In transferring textual information about companies' business model into this vector space, cosine similarity has been proposed to measure how 'closely' a company is oriented towards a particular technology. The overall framework to derive the firm-level indicator of technology usage based on textual innovation data is displayed in Figure 3.

For the proposed measure of technological proximity to be useful, it should satisfy two properties:

(i) It should allow for a differentiation of cleantech oriented firms form non-cleantech oriented firms, i.e. a company whose business model is unrelated to clean technologies should be distant from any of the clean technology embeddings.

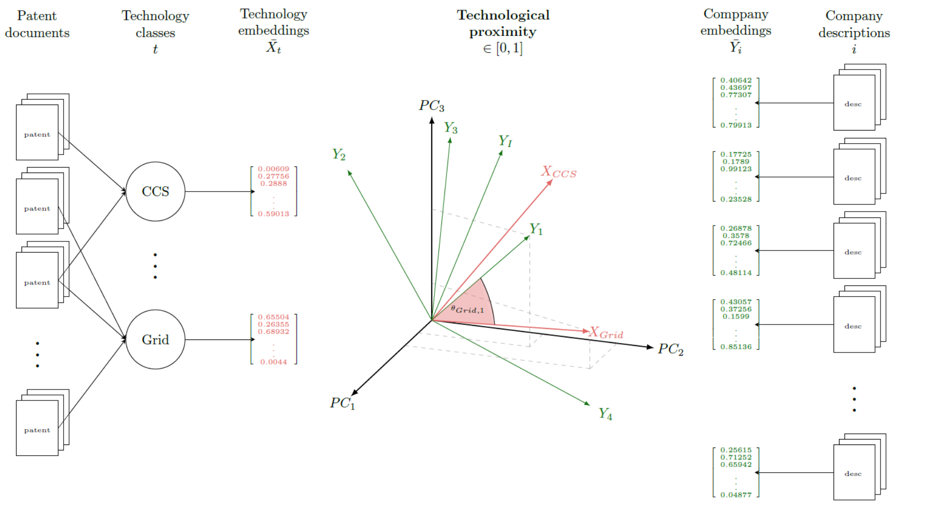
(ii) It should position cleantech companies closest to their most relevant technologies, i.e., a company specialized in geothermal energy should be identified by a relatively high proximity to the technology embedding for renewable energy generation, while at the same time it should show a significantly lower proximity to the other embeddings within the cleantech system, e.g., to the embedding of CCS technologies.

To validate these desirable properties, I use a sample of detailed business summaries of both cleantech and non-cleantech firms. More precisely, the sample comprises business descriptions of all firms that have been listed on the Cleantech 100 list in recent years.[18] They are contrasted

[17] By definition, cosine between two real-valued vectors, which is the case for word embedding based vectors, can take on negative values. Conceptually, this would indicate the the embeddings consist of contextually opposing words. For the purpose of measuring a firms proximity to a technology, it is sufficient to assess how 'closely' the firm is technologically oriented towards a certain technology. A value close to 1 indicates 'high technological proximity' and a value of 0 reflects 'technologically unrelated'. Thus, I truncate negative cosine values to 0.

[18] The Cleantech 100 list is published each year by the Cleantech Group and comprises 100 leading companies in the various clean technology sectors. The list results from an elaborate selection process conducted by an independent expert panel. Starting from an extended nomination list of more than 10,000 firms from close to 100

Figure 3: Illustration of framework to map technologies to company descriptions



Note: For illustration purposes, 384-dimensional technology and company embeddings are displayed on their three principal components (PC).

against business summaries of all companies listed on the S&P 500 constituting the observations of non-cleantech firms.[19] Overall, the sample comprises 533 business summaries of companies that have been listed on the Cleantech 100 list since 2009[20] and business summaries for all of the S&P 500 firms.

It is reasonable to assume that the business models of the firms that make it onto the Cleantech 100 list are closely related to at least one of the clean technology fields derived in Section 3, as an elaborate selection process was conducted to derive the final list. Thus, it is expected that their company embeddings show a relatively high proximity to at least one of the clean technology embeddings, thereby allowing to identify them as cleantech firms. Company embeddings of S&P 500 firms, in contrast, are expected to be more distant from the clean technology embeddings. Following this line of argumentation, the business summaries of the firms on the Cleantech 100 list are labeled as 'cleantech' and business summaries from the S&P 500 firms are labeled as 'non-cleantech'.[21] Based on this labeled dataset, the technological proximity measure is used to classify whether a firm's business model is cleantech oriented or not. This allows to get a first evaluation of the measure's quality, since it should yield low proximity values for any of the non-cleantech firms, while at the same time it should detect

distinct countries, the panel applies objective criteria to derive the final list (Cleantech Group 2022). Business summaries for these cleantech firms are retrieved from https://i3connect.com

[19] Business summaries retrieved from https://www.cnbc.com

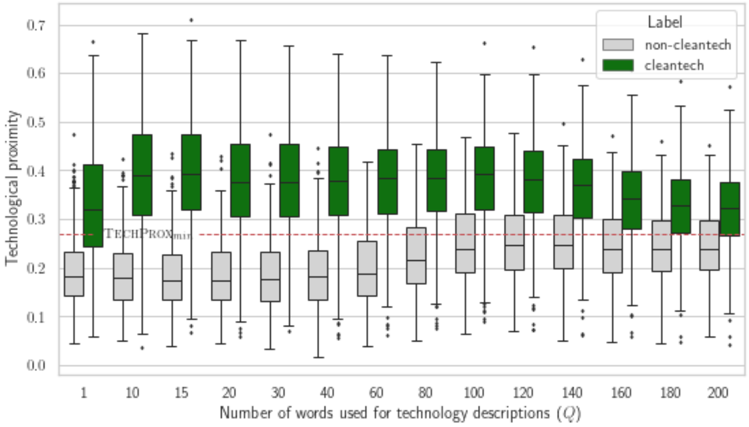
[20] There are several companies that have made it on the Cleantech 100 list in several years, which explains why the total number is not larger.

[21] List of S&P 500 firms has been cleaned by three companies that have also made it onto the Cleantech 100 list. Moreover, after careful validation of the S&P 500 companies' websites, 27 firms which have a clear focus or a major business segment in any of the 10 clean technology fields were labeled as 'cleantech' instead of 'non-cleantech'.

cleantech firms by a high proximity value for their most relevant technology.

Moreover, the binary classification task forms the basis to find the 'optimal' number of words used to model semantic technology descriptions, Q, along with the minimum threshold, TechProx$_{min}$, that must be exceeded for the company to be classified as 'cleantech'. For this purpose, the proximity to all 10 cleantech areas is calculated for each company in the labeled sample and their maximum value, i.e. the proximity value of the firm's most relevant technology, is retained. This step is repeated for different numbers of Q. Figure 4 shows the distribution of TechProx for both the cleantech labeled and non-cleantech labeled companies along different values of Q. The figure suggests that the discriminative 'quality' of the technology proximity measure depends on the number of words, Q, that are used to model the semantic technology description. The more words are used, the worse is the segregation into cleantech and non-cleantech firms. Intuitively, as the number of words increases, terms are added to the technology description that are less relevant in describing the technology, making the description increasingly fuzzy. On the other hand, with an insufficient number of words, the word sequence contains too little context to adequately represent a complex construct such as a technology.

Figure 4: Distinguishing cleantech firms from non-cleantech firms via TechProx



Note: Distribution of technological proximity values between cleantech labeled and non-cleantech labeled firms (for each firm only the highest technological proximity value to the 10 clean technologies is retained, i.e. the proximity value of the technology that is most relevant to the company in semantic vector space). Distribution is displayed as boxplots (median as bar, interquartile range (IQR) as box, 1.5*IQR past the low quartile as lower whisker and 1.5*IQR past the high quartile as upper whisker, values beyond the whiskers as individual points). Distribution is shown for different values of Q, i.e., for different numbers of technical terms used to model technology descriptions. Figure suggests that discriminative power depends on the number of words used to model technologies semantically. With an increasing number of words, terms are added that relate with lower probability to a technology causing the technology description to become fuzzy which diminishes the measure's discriminative power. On the other hand, too few words means that the word sequence contains too little context to adequately represent a complex construct like a technology.

In order to find the optimal values for Q and TechProx$_{min}$, the labeled sample of business summaries is randomly split into a 50% validation set and a 50% test set. On the validation set,

Table 1: Performance of TechProx in distinguishing cleantech from non-cleantech firms

Label	Precision	Recall	F1-Score	Support
Cleantech	0.87	0.86	0.86	284
Non-cleantech	0.83	0.84	0.83	233
			0.85	517

Note: Performance measured on random test set with optimal values of $Q = 15$ and $\text{TechProx}_{min} = 0.27$. Optimal values for Q and TechProx_{min} have been determined on the validation set by tuning F1-Score.

grid search is used to find the optimal values of both parameters. Tuning the F1-Score on the validation set yields an optimal value of $Q = 15$. The optimal value of TechProx_{min} is 0.27. Thus, if TechProx exceeds a value of 0.27, the respective technology is being considered as relevant to the business model of the focal company. In this way, the company is identified as cleantech firm. Given the optimal hyperparameter values found on the validation set, the test set is then used to evaluate the proximity measure's performance in distinguishing cleantech firms from non-cleantech firms. The classification performance metrics are displayed in Table 1. Results show that if the proximity measure detects a firm as a clean technology company, it is correct in almost 9 out of 10 cases, as it can be seen by the 87% precision for the cleantech class. The framework retrieves 86% of all cleantech firms and 84% of all non-cleantech firms in the test dataset (recall). The overall F1-Score is 85%. It is noteworthy that the classification has only been conducted by means of the technology mapping framework that solely relies on business descriptions. Arguably, with additional characteristics such as industry affiliation and patent activities (if applicable), training a classification model could probably improve the identification. These promising results suggest that the proximity measure's first property is satisfied: it allows for an effective discrimination between cleantech and non-cleantech ventures.

Next, I validate the measures capability to position cleantech firms close to their most relevant technology while showing significantly lower proximity to all other technologies within the technological system. To validate this property, I conduct a one-sided Wilcoxon signed-rank test (Wilcoxon 1945). The test allows for a pair-wise comparison of a firm's proximity value of the closest technology with the proximity value of the second closest technology. For each of the clean technology fields, this test is performed within the top 1% (5%) group of companies which show the highest proximity values to the focal technology. In this way, it is tested whether the proximity of a company's most relevant technology is significantly larger than the proximity to the second closest technology. For a disambiguation across clean technology fields this is a desirable property which TechProx is supposed to fulfill. As a further objective measure, I also report the fraction of firms within the top 1% (5%) that originates from the Cleantech 100 list. If the proposed approach provides a reasonable mapping of clean technologies to business models, this fraction is expected to be high, given the technology specialization of the firms on

Table 2: Performance of TechProx in positioning cleantech firms within clean technology space

Clean technology field	**Confidence levels Wilcoxon signed rank test in**		**Fraction of cleantech labeled firms in**	
	top 1%	top 5%	top 1%	top 5%
Adaption	***		1.00	0.87
Battery	**		1.00	0.98
Biofuels	***		1.00	0.96
CCS	**		1.00	0.98
E-Efficiency	**		1.00	1.00
Generation	***	***	1.00	1.00
Grid	***	***	1.00	1.00
Materials			1.00	0.96
Mobility			1.00	0.90
Water	***	***	1.00	0.98

Note: Table reports confidence levels for rejecting the null hypothesis of one-sided Wilcoxon signed rank test for pair-wise comparison of highest TechProx value with second highest TechProx value. Null hypothesis states that the paired differences between a firm's highest TechProx value and second highest TechProx is zero. Tests are based on the top 1% (5%) group of firms with highest proximity to the respective cleantech field. Moreover, table shows fraction of cleantech labeled firms in the top 1% (5%) group of companies with highest proximity to the focal cleantech field. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

the Cleantech 100 list.

Table 2 reports both statistics for each of the clean technology fields. The table shows that the companies on the Cleantech 100 list have the highest proximity in all clean technology areas. Among the top 1%, all companies originate from the Cleantech 100 list, among the top 5% this is still true for at least 87%. For most of the clean technology fields, the technology mapping also allows for a clear demarcation from other clean technology fields. This can be seen in the high confidence levels of rejecting the null hypotheses that the paired differences between a firm's highest TechProx value and second highest TechProx is zero. Only the cleantech areas 'Mobility' and 'Materials' are exempted and show a high proximity to other cleantech areas, impeding a clear-cut technology attribution. Generally, the demarcation diminishes in the group of the top 5% of firms. However, all in all, and after careful validation of the top 1% of companies with the highest proximity values for all of the cleantech fields, it is concluded that the measure performs well in assigning the most relevant clean technology field to cleantech oriented firms. To support this conclusion, Table 3 shows, as an example, the business summaries of the 1% of companies with closest proximity to CCS technologies.

Based on the proposed measurement approach and its desirable properties, the following section identifies technology-oriented entrants within a representative sample of German start-ups. Using survey responses about the start-ups' environmental performance, the section shows distinguishable characteristics of cleantech companies in terms of their ability to act as accelerators of technological change towards decarbonization and dematerialization.

Table 3: Top 1% companies closest to Carbon Capture and Storage (CCS) technology embedding

Business summary	TechProx
Developer of direct air capture technology that safely and permanently removes carbon dioxide from the air ...	0.603
Developer of technologies for the capture of carbon dioxide from the atmosphere at industrial scale ...	0.583
Developer of CO2 capture technology that significantly reduces the costs and environmental impacts of CO2 separation ...	0.567
Developer of energy- and capital-efficient technology for capturing carbon dioxide from industrial sources ...	0.564
Developer and licensor of process technologies to convert carbon dioxide into high-value major chemicals ...	0.547
Developer of carbon dioxide mineralization technology for industrial use in capturing, converting and sequestering carbon emissions as valuable byproducts ...	0.544
Developer of a carbon capture and reuse technology that transforms abundant waste and low-cost resources into low carbon fuels and chemicals ...	0.518
Designer of nanoporous materials for the gas storage and separation industries ...	0.465
Developer of low-cost building materials from industrial carbon dioxide emissions ...	0.457
Developer of methane conversion technology for creating fuels and chemicals from natural gas ...	0.444

Note: Top 1% of companies which show the highest technological proximity to CCS technologies from the sample of Cleantech 100 firms and S&P 500 companies.

4 Technological proximity mapping of new ventures

In this section, the technology mapping framework is applied to a sample of German start-up firms. For this purpose, the study makes use of the IAB/ZEW Start-up Panel as provided by the Research Data Centre of the Centre for European Economic Research (ZEW-FDZ) (Gottschalk 2013). This unique survey data contains detailed firm-level information covering questions about financials, innovation activities and founder characteristics among other variables. Start-ups from all economic sectors are included in the survey. They are drawn from the Mannheim Enterprise Panel (MUP) which covers the near universe of economically active firms (Bersch et al. 2014) in Germany. For the 2018 wave, specific questions about the environmental impact of start-ups' products and services as well as questions about their environmental innovation activities were included in the survey. This makes the survey wave highly suitable for assessing whether clean technology-oriented entrants have distinguishable characteristics that indicate their role as accelerators of a green technological change. For this purpose, I enrich the survey with the start-ups' business descriptions as published in their founding year.[22] Of the 3,789 firms that responded to the environmental-related questions, business descriptions are available for 3,081 of them. For the remaining start-up companies, their archived websites were retrieved from the Internet Archive[23] at the date closest to their founding date. These historical versions of the start-ups' websites are then searched for sub-pages whose link contains keywords such as 'About

[22] Business descriptions are retrieved from the MUP, whose panel structure allows for retrieving the business descriptions at the time of founding.

[23] https://archive.org/

us', 'Products', 'Services', 'Technology' and 'Solutions' in order to extract the textual content found on these sides as an alternative source for their business descriptions. Overall, the final sample comprises 3,269 start-up firms for which survey responses on the environmental-related questions exist and company descriptions close to their time of founding could be recovered. For these companies, business descriptions are used to calculate their technological proximity to each of the 10 clean technology areas. In Figure 5, the distribution of the proximity values is displayed in the form of box-and-whisker plots. The figure shows that the majority of start-ups in the sample have no technological relation to the corresponding technology field, as indicated by the high distribution mass close to zero. At the same time, for each technology, there are a number of companies that stand out with a high technological proximity to the corresponding technology field. These are displayed as 'outliers' in the boxplot and correspond to firms whose proximity value exceeds the upper whisker in the respective distribution. The business descriptions of these start-ups share a high contextual overlap with the semantic representation of the focal clean technology. If these are indeed ventures whose business model builds on clean technological solutions, it is expected that their products and services have a positive environmental impact. In order to verify whether these are indeed market entrants whose products and services are based on environmentally beneficial technologies, the following section makes use of one of the environment-related survey questions.

4.1 Environmental impact of cleantech entrepreneurs' products and services

In the survey, start-ups' were asked to which extent their products and services have a positive environmental impact for their customers. Positive environmental impacts include emission reductions, improved energy efficiency, and better recyclability among other factors.[24] By virtue of their technological orientation, the products and services of cleantech entrepreneurs are expected to have a significant positive environmental impact. In other words, higher values of TechProx should reflect business models whose products and services have positive environmental outcomes for the ultimate users of these products.

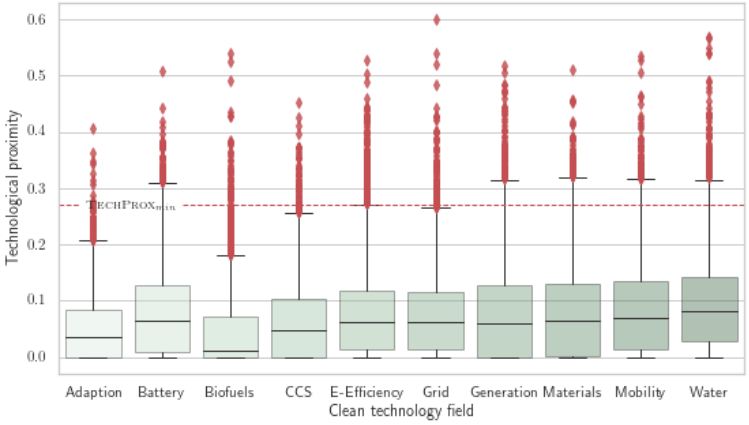
This is tested by regressing the environmental impact of entrants' business models, $EImp$, on TechProx for each of the 10 clean technology fields separately.

$$EImp_i = \beta_0 + \beta_1 \textsc{TechProx}_{t,i} + \boldsymbol{\beta_3 X}_i + \epsilon_i \qquad \forall \;\; t \tag{3}$$

$\boldsymbol{X}$ describes additional firm-level characteristics as control variables. These comprise sector and product type fixed effects which are both expected to capture already some of the variation in the environmental impacts of the firms' products and services. Moreover, it includes variables

[24] See Table 7 in the Appendix for a detailed listing of the environmental impact questions.

Figure 5: TechProx distribution in start-up survey across clean technology fields



Note: Distribution of technological proximity values of start-up firms in the 2018 IAB/ZEW Start-up survey across different clean technology fields. Distribution displayed as boxplots (median as bar, IQR as box, 1.5*IQR past the low quartile as lower whisker and 1.5*IQR past the high quartile as upper whisker, values beyond the whiskers as individual points). Following Tukey (1977), TechProx values exceeding the upper whisker are 'outliers' which correspond to start-ups with a particular high proximity to the respective technology field. Note that the upper whiskers center around the value TechProx$_{min}$ = 0.27 which has been found to discriminate best between cleantech and non-cleantech firms in Section 3.3. This suggests that in this representative sample of German start-up companies, the identification of cleantech ventures via a TechProx value exceeding 0.27 closely matches companies whose proximity value is statistically determined as an outlier. In a representative sample, this seems a desirable property of the measure: it effectively allows for a discrimination of firms whose business model is based on the focal technology from firms whose business model is unrelated to the technology field. With this hard cut-off value, 545 of the 3,269 start-ups are classified as cleantech ventures.

capturing whether the firm conducted R&D, whether it received public support grants and information on the new ventures' financial performance, its size and age as well as information about the founders educational background (see Table 5 for an overview of control variables and their descriptive statistics). Table 4 reports coefficient estimates of the main variable of interest TechProx$_t$.[25] It can been seen that a higher technological orientation towards any of the 10 clean technology fields significantly corresponds with the firms' products and services having a positive environmental impact. Depending on the technology field, a 0.01 increase in TechProx is associated with a 1.2 to 5.0% higher probability of having at least a moderately positive environmental impact compared with no positive impact.

This positive relationship also holds if the start-ups are classified as cleantech or non-cleantech based on the hard cut-off value of TechProx$_{min}$=0.27. This is captured by the variable CleanTech$_t$ which takes on values of 1 if the entrant's technological proximity value exceeds the minimum threshold of 0.27 and 0 otherwise. The significant relationship in this robustness check only vanishes for firms active in technologies for the adaption to climate change and for start-ups providing clean technology solutions in the field of mobility. Overall, the results show that by virtue of their technological orientation, cleantech ventures significantly

[25] Full regression results can be found in Table 8 in the Appendix.

Table 4: Relation between TechProx and the environmental impact of the entrants' products and services, *EImp*

Dependent variable	Clean technology (t)	TechProx$_t$	CleanTech$_t$ (0,1)
EImp	Adaption	1.012*	0.944
	Battery	1.046***	3.083***
	Biofuels	1.049***	1.900**
	CCS	1.050***	2.366**
	E-Efficiency	1.045***	4.319***
	Generation	1.042***	4.375***
	Grid	1.038***	2.156***
	Materials	1.036***	2.234***
	Mobility	1.028***	1.320
	Water	1.034***	2.414***

Note: Environmental impact questions were asked on a Lickert scale with three response possibilities: (1) No positive environmental impact; (2) moderate positive environmental impact; (3) substantial positive environmental impact (see also Table 7 in the Appendix). *EImp* equals (3) substantial positive environmental impact if the firm responded with (3) to at least one of the questions. *EImp* equals (2) moderate positive environmental impact if the firm responded to none of the questions with (3) and to at least one of the questions with (2). Else *EImp* equals (1) no positive environmental impact. Coefficient estimates reported as proportional odds ratios reflecting the factor by which an increase in TechProx$_t$ of one index point (0.01) corresponds to an increase in the odds of having at least a moderate positive environmental impact compared to having no environmental impact (c.p.). Alternatively, coefficient estimates for CleanTech$_t$ reflect by how many times the odds of a start-up classified as cleantech firm in the respective technology field are higher in having at least a moderate positive environmental impact compared to a non-cleantech start-up (c.p.). Estimates correspond to regression model 3 run individually for each technology. Full model results, including coefficient estimates of control variables, can be found in Table 8 in the Appendix. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

contribute to the transition towards decarbonization and dematerialization. Their products and services have a positive impact on their customers' CO$_2$ footprint, lead to a reduced consumption of natural resources and improve recyclability. However, a key research question is whether cleantech entrants also show a higher propensity to introduce additional environmental innovations, i.e., whether, for example, their own R&D efforts lead to a further development of this transformation process. In the following section, I investigate this question by relying on survey information about the firms' environmental innovation activities.

4.2 Environmental innovations among cleantech entrepreneurs

I use a second set of questions that asked firms about their environmental innovation activities to characterize clean technology-focused market participants in terms of their propensity to innovate. Environmental innovations are defined as products and processes which allow the venture to reduce its energy and material consumption or its emissions or to improve the recyclability and durability of its own products.[26] To test whether cleantech entrants, besides their sustainability oriented business models, are additionally characterized by a higher propensity to introduce environmental innovations, I estimate the following regression model.

$$EInno_i = \beta_0 + \beta_1 \text{TechProx}_i + \boldsymbol{\beta_3 X}_i + \epsilon_i \tag{4}$$

[26] See Table 7 in the Appendix for a detailed listing of the environmental innovation questions.

Table 5: Descriptive statistics regression variables

Variable	Description	Mean	SD	Min	Max
TechProx	Degree of start-ups technological proximity to its most relevant technology (i.e. firms highest technological proximity value across the 10 clean technology fields).	0.174	0.098	0	0.599
CleanTech	Indicating whether start-up is classified as cleantech firm or as non-cleantech firm. Cleantech if TechProx exceeds threshold of 0.27.	0.167	0.373		
size	Size of the start-up in number of total employees.	6.330	12.100	1	407
age	Age of start-up in years.	3.000	1.560	1	6
R&D	Indicating whether start-up conducted own research and development in 2017.	0.311	0.463		
R&D intensity	R&D intensity in 2017 measured as number of employees (including founders) which spent at least 50% of their working hours on R&D relative to the total number of employees.	0.106	0.255	0	1
returns	Indicating whether the start-up generated returns in 2017.	0.959	0.198		
break even	Indicating whether the start-up was profitable in 2017.	0.793	0.405		
subsidy	Indicating whether the firm received a public grant in 2017.	0.139	0.346		
team-size	Total number of founders.	1.460	0.809	1	15
university	Indicating whether at least one of the founders holds a university degree.	0.393	0.489		

Note: Table shows descriptive statistics of main variables of interests, TechProx and CleanTech respectively, in regression model 4 as well as for control variables used in regression models 3 and 4. Regression models also include sector fixed effects and product type fixed effects. The latter controls for the following categories: manufacturing of product, service, trade, construction, repair, rental.

The dependent variable, *EInno*, indicates whether or not the venture introduced an environmental innovation after its foundation. The main independent variable of interest, TechProx, refers to the firm's highest technological proximity value across the 10 clean technology fields.

Table 6 reports the average marginal effect estimates for different model specifications. In the most parsimonious specification (1), *EInno* is only regressed on TechProx (CleanTech) controlling for basic firm characteristics such as size and age as well as sector fixed effects. The regression results suggest that, on average, a higher orientation towards clean technologies is associated with a significantly higher probability to introduce environmental innovations. More precisely, cleantech firms' probability to introduce an eco-innovation is, on average, almost 7 percentage points higher as compared to non-cleantech firms. This relationship appears to be highly robust against the inclusion of a wide range of control variates. In model specification (2), for example, innovation-related information are included as additional controls. These comprise an indicator that reflects whether the start-up received a public subsidy, which usually indicates that it is an innovative market entrant. Furthermore, it includes information whether the start-up conducted R&D in 2017 as well as the start-up's R&D intensity, measured as the fraction of employees actively engaged in R&D activities. While the estimates for TechProx (CleanTech) remain unchanged, subsidy recipients and R&D oriented entrants show a significantly higher probability of adopting environmental innovations. Regression specification (3) adds information on the entrants' financial performance which positively correlate with the firms' propensity to eco-innovate. Again, estimates for TechProx (CleanTech) remain robust against inclusion of financial controls. In specification (4), founder characteristics reflecting the absolute number of founders and whether at least one of the founders holds a university degree are added to the regression. Interestingly, the propensity to introduce environmental innovations is significantly lower for firms led by founders with a university degree. Arguably, founders with a more practical educational background, such as craftsmen, are more likely to develop business ideas in which technical environmental innovations are of greater importance. The estimates of the main variables of interest TechProx and CleanTech remain largely robust. Ultimately, specification (5) adds product type fixed effects which control for the start-ups' main type of product or service (manufacturing of product, service, trade, construction, repair, rental). In this final model specification, cleantech firms' probability to introduce environmental innovations is, on average, 7.8 percentage points higher as compared to non-cleantech firms which clearly characterizes them as environmental innovators.

Following the Oslo Manual, a product innovation is defined as 'a product whose technological characteristics or intended uses differ significantly from those of previously produced products' (OECD/Eurostat 2018, p. 32) and a process innovation refers to an 'adoption of technologically new or significantly improved production methods' (OECD/Eurostat 2018, p. 32). Hence, if a venture introduces an environmental product or process innovation, it means that it adapts its products or processes in such a way that they are environmentally superior compared to its previous products and processes. For the case of clean technology-oriented market entrants, the introduction of environmental innovations imply an additional contribution to the diffusion of higher sustainable market standards. Besides their clean technology-oriented business model, they are also characterized by a higher propensity to introduce products and processes that further add to higher environmental standards. Although the results in this section are only of descriptive nature, they suggest that market entrants with a strong focus on clean technological solutions act as accelerators of a technological transition towards green market standards. The distinguishable characteristics of cleantech entrants are in line with entrepreneurship theory that attributes new ventures a special role in this technological transition. While disruptive technological change is barely driven by established firms due to their technological path dependence, new entrants that focus on clean technology solutions are unconstrained to introduce additional and often more radical technology innovations. This gives cleantech entrants a special role as enablers of new technological pathways for sustainable market solutions. The characteristics of cleantech oriented business ventures found in this section support the attribution of this special role. Together with the proposed framework for identifying cleantech companies, this opens a new avenue for entrepreneurship research to demonstrate why cleantech entrepreneurs should be at the center of policies to accelerate the transition to a low-carbon economy.

Table 6: Relation between TechProx and entrants' environmental innovation activity $EInno$

	$EInno$				
	(1)	(2)	(3)	(4)	(5)
TechProx	0.003***	0.003***	0.003***	0.003***	0.003***
log(size)	0.017***	0.013***	0.012***	0.017***	0.015***
age	0.001	0.003	0.001	0.001	0.001
subsidy		0.067**	0.073***	0.080***	0.089***
R&D		0.078***	0.079***	0.105***	0.110***
R&D intensity		−0.055	−0.017	−0.020	−0.040
returns			0.125***	0.110**	0.102**
break even			0.078***	0.065***	0.071***
team size				−0.020*	−0.023*
university				−0.121***	−0.115***
Sector controls	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	Y
N	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.033	0.038	0.043	0.054	0.062
CleanTech	0.068***	0.068***	0.064**	0.060**	0.078***
log(size)	0.018***	0.013***	0.012***	0.017***	0.015***
age	0.001	0.003	0.000	0.001	0.001
subsidy		0.067***	0.074***	0.081***	0.089***
R&D		0.081***	0.082***	0.108***	0.114***
R&D intensity		−0.055	−0.016	−0.020	−0.039
returns			0.126***	0.111**	0.103**
break even			0.078***	0.065***	0.071***
team size				−0.019*	−0.023*
university				−0.122***	−0.115***
Sector controls	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	Y
N	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.033	0.037	0.043	0.054	0.061

Note: Environmental innovation questions were asked on a Lickert scale with three response possibilities: (1) No environmental innovation; (2) environmental innovation with moderate environmental effect; (3) environmental innovation with substantial environmental effect (see also Table 7 in the Appendix). To facilitate interpretation, the response variable was converted to a dichotomous variable, and model 4 was estimated as a logistic regression. Firm is identified as innovator if it responded with at least (2) to at least one of the questions ($EInno = 1$). Else $EInno$ equals 0. Coefficient estimates reported as average marginal effects reflecting the percentage point change in the probability to introduce an environmental innovation if the explanatory variable increases by one unit. Table 9 in the Appendix shows coefficient estimates if ordinal scale of response variable is kept. Results are robust with respect to how the response variable is defined. Change in observation numbers due to item non-response. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

5 Discussion and conclusion

Current research not only suggests that increased investment in advanced low-carbon technologies allows for a further decrease of reduction costs of future emissions (Bistline & Blanford 2020) but also that many near-commercial technologies with substantial emission reduction potential already exist (Bataille et al. 2018). However, additional innovation and policy prioritization with a dedicated mix of policy instruments is required to accelerate the technological transition towards a deep industrial decarbonization (Bataille et al. 2018) and higher sustainability standards (Edmondson et al. 2019). Path dependence in incumbent technology regimes and market externalities for environmental innovations are two economic explanations that justify a policy-induced, directed technical change towards a desirable long-term equilibrium of green growth. In light of technological path dependencies, policymakers are, however, well advised to refine their instruments with respect to companies' willingness to introduce sustainable innovations. Constrained by past technological investments, incumbent firms are typically locked into path-dependent trajectories of their existing technology portfolio with little incentive to stimulate disruptive environmental innovations. New ventures, in contrast, are technologically unconstrained in their innovation decisions, seizing regulatory push and market pull effects for sustainable market solutions with more disruptive innovations (Hockerts & Wüstenhagen 2010). This gives rise to new market entrants as enablers of a green technological transition. Following this theoretical consideration, this study has focused on entrepreneurs whose business models build on clean technology solutions such as renewables, carbon capture and storage or clean water solutions. It is shown that clean technology-oriented market entrants have distinguishable characteristics that indeed suggest that they have an important role to play in the technological transition to higher levels of sustainability. Both by virtue of their business models that build on clean technology solutions as well as by a high propensity to adopt additional environmental innovations, they may act as as accelerators in the transition to more sustainable forms of production, consumption, mobility and housing. This motivates why policymakers should pay special attention to clean technology-oriented market entrants for the design of optimal environmental policy.

First and foremost, policymakers need to know and understand both the technological areas where entrepreneurial activity takes place and the environmental challenges where little entrepreneurship is conducted. While for incumbent firms detailed information through R&D investments and patenting activities allow for assessment of their contributions to the diffusion of sustainable technologies, data availability concerning new ventures is generally limited. In fact, assessing whether a new market entrant bears potential to contribute to the diffusion of clean technology solutions is fundamentally a measurement problem: at the time of founding, innovation-related data to identify an entrant's technological orientation is scarce or even non-existent. This is where the study's main contribution comes into play. With the technology mapping framework presented in this study, it is possible to assess the technological orientation of new ventures at or close to the time of business registration. For this purpose, the framework leverages observable business summaries that new ventures are obliged to report upon registration. Transferring new entrants' business descriptions into technology space by means of state-of-the-art methods from the field of NLP, it is shown that entrants' technological orientation can be determined at a fine granular level of distinct technologies. On an aggregate level, this gives policymakers a first idea to what extent and in which technological areas entrepreneurs are active in the development and diffusion of clean market solutions. Moreover, in the context of directed technical change, the framework provides a useful policy tool. Once a new venture registers, the proposed framework makes it possible to measure the ventures' technological orientation. In this way, policymakers can use the framework to systematically scan business registries for clean technology-focused entrepreneurs. This can be an effective way to direct subsidies to companies with high potential to accelerate green technological change or to pre-select potential candidates for government venture capital funding or public incubator programs.

The framework also opens up new gateways for economic research, particularly by providing a codified approach for identifying cleantech start-ups. Future research can benefit from this, especially for empirical assessments of start-ups' role in overcoming sustainability inertia among path-dependent incumbents. For this purpose, it requires empirical strategies that take a closer look at the interactions between cleantech start-ups and carbon-intensive incumbents. Different channels of innovation interaction exist that deserve closer investigation. In an alliance perspective on environmental innovation activities, established companies may act as source of funding for sustainable entrepreneurs. Besides a high willingness of new ventures to seize market opportunities of green growth by introducing radical environmental innovations, they typically lack capital to scale such innovations. In search for funding, corporate venture capital can be beneficial not only for the new venture but also for the corporate investor. It provides the corporate investor with a source for proof of concepts and allows for experimental learning which requires the investment target to have a certain distance from the investor's accumulated knowledge base (Hegeman & Sørheim 2021). At the same time, the incumbent does not need to leave its existing business model and technology pathway but has some degree of control over the technological advancements which are developed outside its own organization. Once the new technology is mature enough, the incumbent may decide to integrate it as complementary process or product line. In this alliance perspective, the funding of clenteach entrepreneurs through established companies is not just beneficial for both parties but, more importantly, also leads to advances in the transition to more sustainable forms of technology.

There is also a trading perspective in the green technological transition through innovation interactions between incumbents and new ventures. Under increased regulatory pressure, incumbent firms possibly see the need to innovate and adapt their business models more directly. This may incentivize them to pay license fees for the use of clean technologies developed by cleantech start-ups. It may even lead to the acquisition of cleantech start-ups by the regulated incumbent. In this scenario, incumbents would not make risky R&D investments themselves, but could continue to amortize their existing technology investments internally while beginning to build separate product and service lines based on the acquired clean technology solutions. This trading perspective on innovation interactions may yet again be an important channel of accelerating the green technological transition.

Ultimately, there is a competition perspective in overcoming sustainability inertia among incumbents. In the search for new markets and market share, disruptive innovations from cleantech start-ups can force established companies to adapt their existing business model with more radical sustainability innovations. In this way, incumbents may try to preempt future competition in its main product market. Despite their technological path dependence, they may feel forced to respond to increased competition with the introduction of own environmental innovations that eventually disrupt their existing knowledge base. However, this competition perspective may also result in incumbents acquiring entrants to terminate their innovative projects. Established firms may use their financial power to hamper nascent technologies to diffuse as they see their market position threatened by higher sustainability standards. This has been documented before in the pharmaceutical industry, where incumbents terminated innovative projects in the companies they acquired in order to retain their monopoly rents from established technologies (Cunningham et al. 2021).

Presumably, all of these interaction dynamics are technology-specific and industry-dependent. Fundamental to any empirical investigation of these interaction channels is a codified approach to identify cleantech start-ups, preferably at a fine level of distinct technology solutions. Future research could develop empirical strategies to examine these interaction effects and use the framework presented in this paper to identify relevant cleantech entrepreneurs in the first place.

There are limitations to the study. The distinguishable characteristics of cleantech entrants favoring a green technological change have been found by contrasting cleantech start-ups against non-cleantech start-ups. Theory suggests a special role for new entrants because, unlike incumbents, they are not characterized by technological path dependence. Therefore, it would be more desirable to empirically determine entrants' environmental characteristics by contrasting cleantech ventures against incumbents. Unfortunately, the author does not have survey data that includes environmental information on both new and established companies. Furthermore, the technology mapping framework has been applied to company summaries, which can be brief and arguably provide little insight into a company's technology usage. While this can theoretically lead to false negatives in detecting companies that are relevant in a particular technology area, text embedding models alleviate this concern to some extent. This is because they do not depend on exact word matches but place words in vector spaces signaling whether distinct words are close in semantic meaning or not. So even if a business description does not contain technology-specific words, it allows the description's words to be placed into the developed technology space capturing associative meaning between business model and technology. Moreover, the proposed framework has the advantage that it can be applied to any source of textual information about companies. Besides business summaries from business registries, corporate website content poses another promising source of textual data to conduct the technology mapping. I leave it to future research to show how useful webdata is in the mapping of technologies to business models.

References

Acemoglu, D., Aghion, P., Bursztyn, L., & Hemous, D. (2012). The Environment and Directed Technical Change. *American Economic Review*, *102*(1), 131–166. https://doi.org/10.1257/aer.102.1.131

Acs, Z. J., & Audretsch, D. B. (2005). *Entrepreneurship, Innovation and Technological Change*. Now Publishers Inc.

Aghion, P., Dechezleprêtre, A., Hémous, D., Martin, R., & van Reenen, J. (2016). Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. *Journal of Political Economy*, *124*(1), 1–51. https://doi.org/10.1086/684581

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy*, *45*(1), 81–96. https://doi.org/10.1016/j.respol.2015.08.001

Archibugi, D., & Pianta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–468. https://doi.org/10.1016/0166-4972(96)00031-4

Audretsch, D. B., Bozeman, B., Combs, K. L., Feldman, M., Link, A. N., Siegel, D. S., Stephan, P., Tassey, G., & Wessner, C. (2002). The economics of science and technology. *Journal of Technology Transfer*, *27*, 155–203. https://doi.org/10.1023/A:1014382532639

Bataille, C., Åhman, M., Neuhoff, K., Nilsson, L. J., Fischedick, M., Lechtenböhmer, S., Solano-Rodriquez, B., Denis-Ryan, A., Stiebert, S., Waisman, H., Sartor, O., & Rahbar, S. (2018). A review of technology and policy deep decarbonization pathway options for making energy-intensive industry production consistent with the Paris Agreement. *Journal of Cleaner Production*, *187*, 960–973. https://doi.org/10.1016/j.jclepro.2018.03.107

Benner, M. J. (2009). Dynamic or static capabilities? Process management practices and response to technological change. *Journal of Product Innovation Management*, *26*(5), 473–486. https://doi.org/10.1111/j.1540-5885.2009.00675.x

Bersch, J., Gottschalk, S., Mueller, B., & Niefert, M. (2014). The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. *ZEW Discussion Paper*, *14-104*. https://doi.org/10.2139/ssrn.2548385

Bistline, J. E., & Blanford, G. J. (2020). Value of technology in the U.S. electric power sector: Impacts of full portfolios and technological change on the costs of meeting decarbonization goals. *Energy Economics*, *86*, 104694. https://doi.org/10.1016/j.eneco.2020.104694

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1145/2133806.2133826

Bohnsack, R., Pinkse, J., & Kolk, A. (2014). Business models for sustainable technologies: Exploring business model evolution in the case of electric vehicles. *Research Policy*, *43*(2), 284–300. https://doi.org/10.1016/j.respol.2013.10.014

Calel, R., & Dechezleprêtre, A. (2016). Environmental policy and directed technological change: Evidence from the european carbon market. *Review of Economics and Statistics*, *98*(1), 173–191. https://doi.org/10.1162/REST_a_00470

Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity — A Survey. *ACM Computing Surveys*, *54*(2), 1–37. https://doi.org/https://doi.org/10.1145/3440755

Chen, H., Zhang, G., Zhu, D., & Lu, J. (2017). Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technological Forecasting and Social Change*, *119*, 39–52. https://doi.org/10.1016/j.techfore.2017.03.009

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. Y. (2013). An SAO-based text-mining approach for technology roadmapping using patent information. *R&D Management*, *43*(1), 52–74. https://doi.org/10.1111/j.1467-9310.2012.00702.x

Cleantech Group. (2022). Global cleantech 100. http://info.cleantech.com/2019-Global-Cleantech-100-Report-Download.html. Accessed March 3, 2022.

Cohen, B., & Winn, M. I. (2007). Market imperfections, opportunity and sustainable entrepreneurship. *Journal of Business Venturing*, *22*(1), 29–49. https://doi.org/10.1016/j.jbusvent.2004.12.001

Cojoianu, T. F., Clark, G. L., Hoepner, A. G., Veneri, P., & Wójcik, D. (2020). Entrepreneurs for a low carbon world: How environmental knowledge and policy shape the creation and financing of green start-ups. *Research Policy*, *49*(6), 103988. https://doi.org/10.1016/j.respol.2020.103988

Comin, D., Cruz, M., Cirera, X., & Lee, K. M. (2020). Technology within and across firms. *NBER Working Paper*, *28080*. https://doi.org/10.3386/w28080

Cunningham, C., Ederer, F., & Ma, S. (2021). Killer Acquisitions. *Journal of Political Econony*, *129*(3), 649–702. https://doi.org/https://doi.org/10.1086/712506

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Diekhof, J. (2015). Do entrants increase incumbents' innovation activity? Escaping the lock-in, stimulating technological change and the transition towards environmentally friendly vehicles. *Journal of Innovation Economics & Management*, *2015/1*(16), 101–137. https://doi.org/10.3917/jie.016.0101

Dijk, M., Wells, P., & Kemp, R. (2016). Will the momentum of the electric car last? Testing an hypothesis on disruptive innovation. *Technological Forecasting and Social Change*, *105*, 77–88. https://doi.org/10.1016/j.techfore.2016.01.013

Doblinger, C., Surana, K., & Anadon, L. D. (2019). Governments as partners: The role of alliances in U.S. cleantech startup innovation. *Research Policy*, *48*(6), 1458–1475. https://doi.org/10.1016/j.respol.2019.02.006

Edmondson, D. L., Kern, F., & Rogge, K. S. (2019). The co-evolution of policy mixes and socio-technical systems: Towards a conceptual framework of policy mix feedback in sustainability transitions. *Research Policy*, *48*(10), 103555. https://doi.org/10.1016/j.respol.2018.03.010

Erickson, P., Kartha, S., Lazarus, M., & Tempest, K. (2015). Assessing carbon lock-in. *Environmental Research Letters*, *10*(8), 084023. https://doi.org/10.1088/1748-9326/10/8/084023

European Patent Office. (2020). Cooperative Patent Classification. https://www.cooperativepatentclassification.org/home. Accessed September 30, 2020.

Fichter, K., & Clausen, J. (2013). *Erfolg und Scheitern "grüner" Innovationen*. Metropolis.

Gast, J., Gundolf, K., & Cesinger, B. (2017). Doing business in a green way: A systematic review of the ecological sustainability entrepreneurship literature and future research directions. *Journal of Cleaner Production*, *147*, 44–56. https://doi.org/10.1016/j.jclepro.2017.01.065

Geum, Y., Lee, H. J., Lee, Y., & Park, Y. (2015). Development of data-driven technology roadmap considering dependency: An ARM-based technology roadmapping. *Technological Forecasting and Social Change*, *91*, 264–279. https://doi.org/10.1016/j.techfore.2014.03.003

Goldstein, A., Doblinger, C., Baker, E., & Anadón, L. D. (2020). Patenting and business outcomes for cleantech startups funded by the Advanced Research Projects Agency-Energy. *Nature Energy*, *5*, 803–810. https://doi.org/10.1038/s41560-020-00683-8

Gottschalk, S. (2013). The Research Data Centre of the Centre for European Economic Research (ZEW-FDZ). *Journal of Contextual Economics – Schmollers Jahrbuch*, *133*(4), 607–618.

Graham, S. J., & Sichelman, T. (2008). Why Do Start-ups Patent? *Berkeley Technology Law Journal*, *23*(3), 1063–1097.

Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject-action-object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change*, *105*, 27–40. https://doi.org/10.1016/j.techfore.2016.01.028

Hegeman, P. D., & Sørheim, R. (2021). Why do they do it ? Corporate venture capital investments in cleantech startups. *Journal of Cleaner Production*, *294*, 126315. https://doi.org/10.1016/j.jclepro.2021.126315

Helmers, C., & Rogers, M. (2011). Does Patenting Help High-Tech Start-Ups? *Research Policy*, *40*(7), 1016–1027. https://doi.org/10.1016/j.respol.2011.05.003

Henderson, R. (1993). Underinvestment and Incompetence as Responses to Radical Innovation: Evidence from the Photolithographic Alignment Equipment Industry. *The RAND Journal of Economics*, *24*(2), 248. https://doi.org/10.2307/2555761

Hockerts, K., & Wüstenhagen, R. (2010). Greening Goliaths versus emerging Davids - Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship. *Journal of Business Venturing*, *25*(5), 481–492. https://doi.org/10.1016/j.jbusvent.2009.07.005

Horne, J., & Fichter, K. (2022). Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. *Journal of Cleaner Production*, *349*(February 2021), 131163. https://doi.org/10.1016/j.jclepro.2022.131163

Hötte, K. (2020). How to accelerate green technology diffusion? Directed technological change in the presence of coevolving absorptive capacity. *Energy Economics*, *85*. https://doi.org/10.1016/j.eneco.2019.104565

Hwang, S., & Shin, J. (2019). Extending technological trajectories to latest technological changes by overcoming time lags. *Technological Forecasting and Social Change*, *143*(February), 142–153. https://doi.org/10.1016/j.techfore.2019.04.013

Kant, M. (2018). *Sustainable Entrepreneurship for Commercialising Radical Clean Technologies - A Means to Enable Carbon Dioxide Utilisation?* (Doctoral dissertation).

Lee, S., Lee, S., Seol, H., & Park, Y. (2008). Using patent information for designing new product and technology: Keyword based technology roadmapping. *R&D Management*, *38*(2), 169–188. https://doi.org/10.1111/j.1467-9310.2008.00509.x

Leendertse, J., van Rijnsoever, F. J., & Eveleens, C. P. (2021). The sustainable start-up paradox: Predicting the business and climate performance of start-ups. *Business Strategy and the Environment*, *30*(2), 1019–1036. https://doi.org/10.1002/bse.2667

Malen, J., & Marcus, A. A. (2017). Promoting clean energy technology entrepreneurship: The role of external context. *Energy Policy*, *102*(March 2016), 7–15. https://doi.org/10.1016/j.enpol.2016.11.045

Moehrle, M. G., Wustmans, M., & Gerken, J. M. (2018). How business methods accompany technological innovations – a case study using semantic patent analysis and a novel informetric measure. *R&D Management*, *48*(3), 331–342. https://doi.org/10.1111/radm.12307

OECD/Eurostat. (2018). *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation* (4th ed.). OECD Publishing. https://www.oecd.org/science/oslo-manual-2018-9789264304604-en.htm

Patel, P., & Pavitt, K. (1997). The technological competencies of the world's largest firms: Complex and path-dependent, but not much variety. *Research Policy*, *26*(2), 141–156. https://doi.org/10.1016/S0048-7333(97)00005-X

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, (August), 248–256.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Rennings, K. (2000). Redefining innovation - Eco-innovation research and the contribution from ecological economics. *Ecological Economics*, *32*(2), 319–332. https://doi.org/10.1016/S0921-8009(99)00112-3

Schaltegger, S., Lüdeke-Freund, F., & Hansen, E. G. (2016). Business Models for Sustainability: A Co-Evolutionary Analysis of Sustainable Entrepreneurship, Innovation, and Transformation. *Organization and Environment*, *29*(3), 264–289. https://doi.org/10.1177/1086026616633272

Schaltegger, S., & Wagner, M. (2011). Sustainable entrepreneurship and sustainability innovation: Categories and interactions. *Business Strategy and the Environment*, *20*(4), 222–237. https://doi.org/10.1002/bse.682

Sick, N., Nienaber, A. M., Liesenkötter, B., vom Stein, N., Schewe, G., & Leker, J. (2016). The legend about sailing ship effects – Is it true or false? The example of cleaner propulsion technologies diffusion in the automotive industry. *Journal of Cleaner Production*, *137*, 405–413. https://doi.org/10.1016/j.jclepro.2016.07.085

Smink, M. M., Hekkert, M. P., & Negro, S. O. (2015). Keeping sustainable innovation on a leash? Exploring incumbents' institutional strategies. *Business Strategy and the Environment*, *24*(2), 86–101. https://doi.org/10.1002/bse.1808

Song, K., Kim, K. S., & Lee, S. (2017). Discovering new technology opportunities based on patents: Text-mining and F-term analysis. *Technovation*, *60-61*(August 2015), 1–14. https://doi.org/10.1016/j.technovation.2017.03.001

Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, *115*, 131–142. https://doi.org/10.1016/j.techfore.2016.09.028

Trautwein, C. (2021). Sustainability impact assessment of start-ups – Key insights on relevant assessment challenges and approaches based on an inclusive, systematic literature review. *Journal of Cleaner Production*, *281*, 125330. https://doi.org/10.1016/j.jclepro.2020.125330

Tukey, J. W. (1977). *Exploratory Data Analysis by John W. Tukey.* http://www.jstor.org/stable/2529486

United Nations. (2015). Paris Agreement. https://unfccc.int/sites/default/files/english_paris_agreement.pdf. Accessed October 1, 2020.

United Nations Environment Program, & European Patent Office. (2015). Climate change mitigation technologies in Europe - evidence from patent and economic data. https://personal.lse.ac.uk/dechezle/climate_change_mitigation_technologies_europe_en.pdf. Accessed September 30, 2020.

Unruh, G. C. (2000). Understanding carbon lock-in. *Energy Policy*, *30*(4), 317–325. https://doi.org/10.1016/S0301-4215(01)00098-2

Wang, Y., Hou, Y., Che, W., & Liu, T. (2020). From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, *11*(7), 1611–1630. https://doi.org/10.1007/s13042-020-01069-8

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179–191. https://doi.org/10.1016/j.techfore.2016.01.015

Appendix

Figure 6: Descriptive statistics textual data

Source	Number of documents (N)	Document length (number of tokens)				Vocabulary size (V)	Preprocessing steps
		Min	Median	Max	SD		
Patent abstracts	559,367	8	123	2,478	79.38	370,110	lemmatization, remove punctuation, remove digits, lowercasing
Cleantech 100	533	4	14	44	6.74	7,831	-
S&P 500	500	92	155	194	17.40	76,290	-
Start-up Survey	3,269	1	18	292	25.57	82,458	-

Note: Table shows descriptive statistics of the different textual data sources used in this paper. Patent abstracts are drawn from EPO's World Patent Statistical database (PATSTAT). Business summaries of firms on the Cleantech 100 list (https://i3connect.com) and S&P 500 (https://www.cnbc.com) are webscraped. Business summaries of firms in IAB/ZEW Start-up Panel are drawn from the Mannheim Enterprise Panel (MUP).

Figure 7: 2018 IAB/ZEW Start-up survey questions on environmental impacts and environmental innovation

Environmental impact

Does your company offer products or services which have the following environmental effects on the customer or the end user?

1. Reduction of energy consumption or CO_2 footprint for the customer.
2. Reduction of other emissions to the air, water, soil or noise for the the customer.
3. Reduction of material or resource consumption, for instance water, for the customer.
4. Improvement of recyclability of customer's products.
5. Improvement of durability of customer's products.

Environmental innovation

Since its inception, has your company introduced innovations that have impacted the environment as follows?

1. Reduction of energy consumption or the overall CO_2 balance in your company.
2. Reduction of other emissions to the air, water, soil or noise in your company.
3. Reduction of material or resource consumption, for instance water, in your company.
4. Improvement of recyclability of your own products.
5. Improvement of durability of your own products.

Note: The questions have been asked on a Likert response scale with the following response possibilities. (1) No; (2) Yes, somewhat; (3) Yes, substantial.

Table 7: Semantic technology descriptions

Adaption		Battery		Biofuels		CCS		E-Efficiency		Generation		Grid		Materials		Mobility		Water	
term	prob	term	prob	term	prob	term	prob	term	prob	term	prob	term	prob	term	prob	term	prob	term	prob
plant	0.028	fuel	0.045	biogas	0.024	gas	0.032	heat	0.016	wind	0.023	battery	0.039	gas	0.014	exhaust	0.025	water	0.016
nucleic	0.014	cell	0.036	fuel	0.021	absorption	0.016	power	0.016	solar	0.023	energy	0.022	furnace	0.009	engine	0.025	waste	0.014
polypeptide	0.013	gas	0.018	gas	0.018	dioxide	0.014	voltage	0.012	rotor	0.018	cell	0.020	material	0.007	combustion	0.020	sludge	0.010
trait	0.010	membrane	0.013	biomass	0.016	carbon	0.013	circuit	0.012	turbine	0.016	charge	0.017	catalyst	0.007	gas	0.016	material	0.008
acid	0.010	anode	0.011	fermentation	0.015	air	0.010	supply	0.010	blade	0.015	storage	0.016	process	0.006	internal	0.014	fraction	0.006
yield-related	0.010	cathode	0.011	fermenter	0.014	stream	0.010	control	0.008	layer	0.010	electrode	0.011	powder	0.006	air	0.012	wastewater	0.006
expression	0.010	electrode	0.011	reactor	0.010	CO2	0.009	switch	0.008	tower	0.010	electrical	0.009	reactor	0.006	drive	0.008	process	0.006
encode	0.010	electrolyte	0.009	plant	0.007	overspray	0.009	steam	0.008	photovoltaic	0.009	heat	0.009	reaction	0.005	fuel	0.007	tank	0.005
present	0.009	hydrogen	0.008	percolate	0.007	flow	0.008	lamp	0.008	cell	0.008	accumulator	0.009	stream	0.005	flow	0.006	treatment	0.005
protein	0.009	layer	0.008	combustion	0.006	stage	0.007	current	0.008	power	0.008	electrochemical	0.008	heat	0.005	motor	0.006	mixture	0.005
enhance	0.007	stack	0.008	tank	0.006	exhaust	0.007	gas	0.008	energy	0.007	power	0.008	melt	0.005	vehicle	0.006	flotation	0.004
modulate	0.007	catalyst	0.007	pyrolysis	0.006	process	0.007	converter	0.007	generator	0.007	electrolyte	0.008	mixture	0.005	control	0.006	separate	0.004
concern	0.006	reformer	0.007	engine	0.006	mixture	0.007	exchanger	0.007	module	0.006	electric	0.007	temperature	0.005	system	0.006	suspension	0.004
invention	0.006	supply	0.006	methane	0.006	heat	0.006	air	0.007	organic	0.006	vehicle	0.007	product	0.004	catalytic	0.006	basin	0.004
method	0.006	water	0.005	waste	0.005	adsorption	0.006	energy	0.007	plant	0.005	lithium	0.006	step	0.004	torque	0.006	filter	0.004
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note: Table shows top 15 terms that describe each of the 10 clean technology fields with highest probability. Terms are learned empirically from corpus of patent abstracts using L-LDA.

Table 8: Relation between TECHPROX and the environmental impact of the entrants' products and services *EImp* (full model results)

	EImp									
t	Adaption	Battery	Biofuels	CCS	E-Efficiency	Generation	Grid	Materials	Mobility	Water
TECHPROX$_t$	1.012*	1.046***	1.049***	1.050***	1.045***	1.042***	1.038***	1.036***	1.028***	1.034***
log(size)	1.042	1.029	1.026	1.024	1.049	1.050	1.037	1.032	1.040	1.023
age	0.998	0.983	0.988	0.992	0.985	0.985	0.988	0.989	0.994	0.995
R&D	1.850***	1.821***	1.851***	1.838***	1.860***	1.880***	1.835***	1.839***	1.829***	1.834***
R&D intensity	0.867	0.835	0.826	0.852	0.873	0.865	0.843	0.848	0.864	0.860
subsidy	1.399***	1.431***	1.419***	1.398***	1.402***	1.412***	1.416***	1.417***	1.426***	1.384***
returns	1.393*	1.295	1.342	1.329	1.253	1.289	1.379	1.275	1.335	1.373
break even	1.046	1.049	1.053	1.043	1.046	1.048	1.068	1.032	1.026	1.048
team size	0.936	0.923	0.924	0.921	0.921	0.923	0.926	0.928	0.926	0.933
university	0.795***	0.792***	0.807**	0.810**	0.819**	0.792***	0.813**	0.809**	0.810**	0.806**
Sector controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Product type controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774
Pseudo R^2	0.059	0.072	0.069	0.070	0.073	0.064	0.072	0.068	0.068	0.067
CLEANTECH$_t$	0.944	3.083***	1.900**	2.366**	4.319***	4.375***	2.156***	2.234***	1.320	2.414***
log(size)	1.041	1.048	1.041	1.040	1.053	1.045	1.037	1.042	1.041	1.037
age	0.998	0.993	0.997	0.998	0.993	0.993	0.992	0.994	0.997	0.998
R&D	1.874***	1.866***	1.875***	1.863***	1.920***	1.917***	1.852***	1.878***	1.869***	1.855***
R&D intensity	0.867	0.865	0.864	0.866	0.860	0.868	0.864	0.867	0.868	0.870
subsidy	1.404***	1.404***	1.400***	1.390***	1.396***	1.399***	1.417***	1.395***	1.403***	1.383***
returns	1.375	1.358	1.388*	1.359	1.318	1.333	1.392*	1.348	1.382*	1.422*
break even	1.049	1.059	1.051	1.047	1.061	1.065	1.061	1.049	1.046	1.051
team size	0.937	0.927	0.935	0.933	0.922	0.935	0.934	0.930	0.936	0.942
university	0.804**	0.802**	0.805**	0.805**	0.807**	0.786***	0.811**	0.806**	0.806**	0.802**
Sector controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Product type controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774	2,774
Pseudo R^2	0.058	0.063	0.059	0.060	0.069	0.059	0.068	0.061	0.061	0.062

Note: Coefficient estimates reported as proportional odds ratios. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Table 9: Relation between TechProx and entrants' environmental innovation capacity *EInno* (ordered logit)

	EInno					
	(1)	(2)	(3)	(4)	(5)	(6)
TechProx	1.015***	1.014***	1.013***	1.013***	1.012***	1.014***
log(size)		1.190***	1.140***	1.125***	1.186***	1.175***
age		1.001	1.010	1.001	1.005	1.012
subsidy			1.317***	1.353***	1.413***	1.456***
R&D			1.427***	1.434***	1.605***	1.675***
R&D intensity			0.780	0.910	0.904	0.815
returns				1.743***	1.633**	1.551**
break even				1.295***	1.226**	1.237**
team size					0.899**	0.887**
university					0.614***	0.627***
Sector controls	Y	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	N	Y
N	3,269	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.022	0.026	0.030	0.033	0.041	0.047
CleanTech	1.339***	1.328***	1.323***	1.295***	1.287***	1.380***
log(size)		1.192***	1.140***	1.125***	1.186***	1.175***
age		1.000	1.009	1.000	1.004	1.012
subsidy			1.323***	1.358***	1.419***	1.461***
R&D			1.448***	1.453***	1.626***	1.704***
R&D intensity			0.778	0.909	0.902	0.817
returns				1.751***	1.641**	1.563**
break even				1.293***	1.223**	1.235**
team size					0.900**	0.888**
university					0.612***	0.627***
Sector controls	Y	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	N	Y
N	3,269	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.021	0.025	0.029	0.033	0.040	0.047

Note: Environmental innovation questions were asked on a Lickert scale with three response possibilities: (1) No environmental innovation; (2) environmental innovation with moderate environmental effect; (3) environmental innovation with substantial environmental effect (see also Table 7 in the Appendix). *EInno* equals (3) environmental innovation with substantial environmental effect if the firm responded with (3) to at least one of the questions. *EInno* equals (2) if the firm responded to none of the questions with (3) and to at least one of the questions with (2). Else *EInno* equals (1) no environmental innovation. Coefficient estimates reported as proportional odds ratios reflecting the factor by which an increase in TechProx of one index point (0.01) corresponds to an increase in the odds of having introduced a innovation with at a least moderate environmental effect compared to having introduced no environmental innovation (c.p.). Alternatively, coefficient estimates for CleanTech reflect by how many times the odds of a start-up classified as cleantech firm in the respective technology field are higher in having introduced a innovation with at least a moderate environmental effect compared to a non-cleantech start-up (c.p.). Change in observation numbers due to item non-response. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Bibliography

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the Z-Score and ZETA® models. In A. R. Bell, C. Brooks, & M. Prokopczuk (Eds.), *Handbook of research methods and applications in empirical finance* (pp. 428–456). Edward Elgar Publishing. https://doi.org/10.4337/9780857936097.00027

Athey, S., & Imbens, G. W. (2019). Machine learning methods economists should know about. *Annual Review of Economics*, *11*, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Calel, R., & Dechezleprêtre, A. (2016). Environmental policy and directed technological change: Evidence from the european carbon market. *The Review of Economics and Statistics*, *98*(1), 173–191. https://doi.org/10.1162/REST_a_00470

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. https://doi.org/10.48550/arXiv.1911.02116

Dörr, J. O. (2022). *Mapping technologies to business models: An application to clean technologies and entrepreneurship* [Unpublished manuscript].

Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022a). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *PLoS ONE*, *17*(2), e0263898. https://doi.org/10.1371/journal.pone.0263898

Dörr, J. O., Licht, G., & Murmann, S. (2022b). Small firms and the COVID-19 insolvency gap. *Small Business Economics*, *58*, 887–917. https://doi.org/10.1007/s11187-021-00514-4

Einav, L., & Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*, *14*, 1–24. https://doi.org/10.1086/674019

Federal Ministry of Finance. (2020). German stability programme 2020. Retrieved December 28, 2020, from https://www.bundesregierung.de/breg-de/service/publikationen/german-stability-programme-2020-1749510

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, *102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, *124*(5), 1423–1465. https://doi.org/10.1086/688176

Hockerts, K., & Wüstenhagen, R. (2010). Greening Goliaths versus emerging Davids - Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship. *Journal of Business Venturing*, *25*(5), 481–492. https://doi.org/10.1016/j.jbusvent.2009.07.005

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Aomputational Linguistics*, 328–339. https://doi.org/10.48550/arXiv.1801.06146

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Olteanu, Y., & Fichter, K. (2022). Startups as sustainability transformers: A new empirically derived taxonomy and its policy implications. *Business Strategy and the Environment*, 1–17. https://doi.org/10.1002/bse.3065

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256. http://www.aclweb.org/anthology/D09-1026

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Ruder, S., Peters, M., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 15–18. https://doi.org/10.18653/v1/N19-5004

Schumpeter, J. A. (1942). *Capitalsim, Socialism and Democracy*. New York: Harper.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

The Economist. (2021). Enter third-wave economics. Retrieved October 25, 2021, from https://www.economist.com/briefing/2021/10/23/enter-third-wave-economics

The Economist. (2020). The corporate undead: What to do about zombie firms. Retrieved September 26, 2020, from https://www.economist.com/leaders/2020/09/24/what-to-do-about-zombie-firms

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839