poid Programme on Innovation and Diffusion

Working Paper

No.041
October 2022

From public labs to private firms: magnitude and channels of R&D spillovers

Antonin Bergeaud
Arthur Guillouzouic
Emiric Henry
Clément Malgouyres

LSE The London School of Economics and Political Science

UKRI Economic and Social Research Council

Abstract
Introducing a new measure of scientific proximity between private firms and public research groups and exploiting a multi-billion euro financing program of academic clusters in France, we provide causal evidence of spillovers from academic research to private sector firms. Firms in the top quartile of exposure to the funding shock increase their R&D effort by 20% compared to the bottom quartile. We exploit reports produced by funded clusters, complemented by data on labor mobility and R&D public–private partnerships, to provide evidence on the channels for these spillovers. We show that spillovers are driven by outsourcing of R&D activities by the private to the public sectors and, to a lesser extent, by labor mobility from one to the other and by informal contacts. We discuss the policy implications of these findings.

Key words: knowledge spillovers, policy instruments, technological distance.
JEL codes: O32; O38; R12

We are grateful to participants at the NBER Summer Institute, CEPR IMO & EMT conference, as well as seminar participants at LSE, Collège de France, INSEAD, Mines ParisTech, Le Mans, Cergy University, Hitotsubashi University, for insightful comments and suggestions. This work also benefited from discussions with Philippe Aghion, Morten Bennedsen, Shawn Kantor, and Nicolas Serano-Velarde. We acknowledge financial support from the French Ministry of Higher Education and Research for this project. We thank the Agence Nationale de la Recherche for sharing data about the LabEx program.

Antonin Bergeaud, HEC Paris, CEPR and Centre for Economic Performance, London School of Economics. Arthur Guillouzouic, Paris School of Economics. Emeric Henry, Sciences Po and CEPR. Clément Malgouyres, CREST-PSE.

Published by
Programme on Innovation and Diffusion
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

This paper was produced as part of the Programme on Innovation Diffusion (POID) financed by the Economic and Social Research Council.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means without the prior permission in writing of the publisher nor be issued to the public or circulated in any form other than that in which it is published.

Requests for permission to reproduce any article or part of the Working Paper should be sent to the editor at the above address.

© A. Bergeaud, A. Guillouzouic, E. Henry and C. Malgouyres, submitted 2022.

1 Introduction

In December 2020, less than a year after the onset of the Covid-19 pandemic, the first dose of clinically-approved vaccine was administered. This resounding technological success is widely seen as the result of a fruitful interaction between public research and the private sector (Cross et al., 2021; Kiszewski et al., 2021). Cross et al. (2021) show that public and charitable financing accounts for more than 97% of funding for the vaccine technology research underlying the Oxford-AstraZeneca vaccine. Similarly, Kiszewski et al. (2021) argue, in the US context, that "NIH funding contributed substantially to the advance of technologies available for rapid development of COVID-19 vaccines".

The existence and magnitude of spillovers from the public research sector to private firms is a long-standing question (since at least Jaffe, 1989). While Azoulay et al. (2019) show that a $10 million boost in NIH funding leads to a net increase of 2.3 patents in the biotechnology and pharma industries, there is limited causal evidence on the effects of public research funding on firms in other sectors.[1] Moreover, there is a lack of systematic empirical evidence on the channels through which these spillovers operate. In this paper, we shed light on these questions, exploiting a large scale funding program of public research in France, the LabEx (*"Laboratoire d'Excellence"*) program, which allocated 1.5 billion euros to 170 academic clusters potentially linked to many different sectors.

We first propose a new measure of scientific proximity between public research groups and industrial sectors, allowing us to measure the exposure of private firms to the program. Based on this measure, we use the funding shock to estimate the causal impact of public research on private sector outcomes. We find that firms spatially and scientifically "close" to funded research groups increase their spending on R&D inputs and achieve higher levels of R&D outputs compared to less exposed firms. Using the grades obtained by unsuccessful LabEx candidates, we build a number of robustness and Placebo tests which confirm our finding. We then use a wealth of qualitative and quantitative evidence to delve into the mechanisms driving these positive spillovers. We show in particular the importance of contracting between firms and public research groups.

The first step to assess empirically the magnitude of spillovers between the public and private sectors is to measure the proximity of the academic clusters with the local industry, in order to identify the firms most likely to be affected by the funding shock. Our first contribution is to

1. On the contrary, there is now a large body of literature documenting spillovers and the importance of knowledge flows within academia (Waldinger, 2012; Moser, Voena, and Waldinger, 2014; Iaria, Schwarz, and Waldinger, 2018).

construct a new measure of proximity. The key idea is to use the distance between the science used by firms and the science produced by research groups, as a proxy for the share of ideas produced by the research group which are relevant for a firm in a given industry.[2] Importantly, this measure assigns a scientific position to firms rather than using the typical approach in the literature that attempts to assign a technological position to an academic group.[3] An academic group and an industry can be close according to our measure, even if the papers published by the group are not yet cited in patents, and would thus be categorized as distant with the traditional approach.

We then test and exploit this measure in the context of the LabEx program that selected 170 academic clusters in 2010 and 2011. These clusters bring together public researchers from different research units, not necessarily from the same institution, planning to work together on a common theme. The funding was run as a competition, with an international jury evaluating 436 proposals. We obtained a number of key pieces of data from the agency organizing the competition, for both accepted and rejected projects, in particular the full bibliography of the proposals and the grades awarded by the jury. The articles listed in the bibliography of the proposal are used to construct the proximity between LabEx projects and industrial sectors through the above procedure, while the identity and grades of rejected projects are the basis for various robustness checks.

Our second contribution is to use this natural experiment to estimate the causal effect of a positive shock in funding of public sector research on private sector R&D. For a given pair of commuting zone (henceforth "CZ") and industry, we calculate exposure as the sum of funding obtained by the LabEx proposals in the CZ, weighted by the proximity of the LabEx to this industry. We then implement a difference-in-differences estimation that exploits the fact that a given industry in a given commuting zone will be more exposed to the shock if it is close technologically and geographically to a funded LabEx. We show that firms in high exposure pairs of CZ and sector significantly increase their employment in R&D after the start of the program, controlling for CZ specific time effects. The magnitude is large: a firm in the top quartile of exposure increases total spending on wages of R&D workers by more than 20% compared to the bottom quartile. We also find significant impacts on outputs of the R&D process in the more exposed sectors, in particular on the creation of new plants and on the production of new patents.

2. Specifically, our proximity measure between the group and a given industry is the sum over all journals of the product of the share of publications of the research group in that journal and the share of citations to that journal coming from patents obtained by firms in the industry.

3. The academic patents or the patents citing the research are typically used to assign a technological distance. We explain later in the paper why our measure is more adapted in our context.

Our identification relies on the assumption that, absent the policy reform, the trends would have been similar between more and less exposed industries. We rely on two sources of variation, variations due to the selection process and variations linked to differences in exposure of industries across commuting zones. These two sources of variations could both raise some challenges to identification. We conduct a number of robustness checks to address these concerns, exploiting in particular the unique feature of our data, the fact we have information on rejected projects and on the grades obtained by projects.

The first concern is that, even in the presence of exogenous shocks, exposure of the industries might not be random. We follow Borusyak and Hull (2021) and consider counterfactual realizations of the shocks. We use grades and requested amounts to build these alternative scenarios. Overall, we show that actual exposure affects outcomes while the counterfactual exposure does not. The second concern is that the shock itself might not be exogenous under the unlikely scenario that the jury chose academic clusters because of the sectors they might affect, and picked those affecting potentially booming sectors.[4] We show that the results are not affected when we restrict to certain ranges of grades.

Our third main contribution is to shed light on the channels through which the important spillovers from public to private research occur. There are three main candidates. First, spillovers could be due to direct collaboration between researchers in the academic cluster and the exposed firms (see e.g. Fernandes and Ferreira, 2013). Second, they could result from mobility of researchers from the public to the private sector or creation by these researchers of startups (see e.g. Agrawal and Henderson, 2002). Third, spillovers could simply be due to informal contacts between researchers from the public and the private sectors, for instance during events organized by the LabEx (see e.g. Dahl and Pedersen, 2004).

Consider the LabEx called ACTION, located in Dijon, working on the development and integration of smart systems, that received 8 million euros in funding through the LabEx program.[5] By 2015, it had already developed close links with the industry, giving rise to 4 patents, 3 EU projects submitted and 2 start-ups related to the LabEx. The initial report written by the LabEx for the funding agency provides clues on how these spillovers to the private sector materialized.[6] It mentions that the Labex plans to sign contracts with firms for PhD co-supervision

4. We view this as unlikely as the international jury members were not informed of the characteristics of the local industries.

5. The activities of the LabEx are summarized as follows: "The project aims to explore the potential of nanotechnology and computing for developing miniaturized systems with new functionalities for applications in the fields of health, transport, energy. That miniaturization will allow technologies to integrate, for example, sensors interconnected and open to the outside world, computers, software, etc., in order to design so-called "intelligent" systems that adapt and anticipate to better respond to the use made of it."

6. These reports requested by the funding agency to formalize the governance of the clusters, were written in

and joint research, to provide training and seminars for people in the industry and to create a "club of partners", described as a "structure of exchange of information between the members of the LabEx and potential partners".

For the entire set of funded projects, 75% of the reports mention the contracting channel, for instance public-private research partnerships, contracts for co-supervision of PhDs or licensing contracts for patents. 52% of the reports mention the mobility channel, with special focus on helping researchers to create spin-offs and encouraging mobility of master and PhD students. Finally, 30% mention informal contacts as an important element, facilitated by the events organized by many of these academic clusters oriented towards a private sector audience. The reports thus highlight all channels, with a predominant role for the contracting channel.

The evidence in the reports provide a comprehensive view of the mechanisms underlying these spillovers, but do not provide counterfactuals for the non funded proposals. We thus provide additional elements to show causal evidence on some specific dimensions. To document the first channel, we obtained data on a program of co-financing of PhDs by firms and public research institutions, called Cifre, involving explicit contracting. We also use data on scientific sub-contracting by firms. To capture the second channel, we use the mobility of researchers that we observe in the administrative data. Using the same identification strategy as in our main analysis, we find that private firms more exposed to the LabEx program significantly increase the likelihood of signing contracts formalizing PhD co-supervision. More generally, we find that the total amount of contracting between public and private labs rises in the most exposed industries quickly after the shock. Finally, we also show evidence of more frequent movements of workers from the public research sector to private firms. Overall, we find evidence for all three channels, with the contracting channel playing a central role.

Our results suggest that financing public research is a powerful policy instrument to spur private sector R&D. While studying the relative importance of this instrument as opposed to more direct financing instruments such as tax credit policies is beyond the scope of the paper, we nevertheless conclude by giving some elements of comparison. France has an extensive R&D tax credit program ("*Credit Impôt Recherche*" or CIR) which represents above 6 billion euros of fiscal spending per year. We document that the distribution of benefits from the CIR across industries is very different from the distribution of indirect benefits from the LabEx program. We suggest that the indirect instrument might better target research intensive firms, which are the only firms in a position to exploit the findings of the public sector.

2012, very early in the project and correspond to projected channels.

Literature review. Bloom, Van Reenen, and Williams (2019) and Teichgraeber and Van Reenen (2022) survey the existing literature on instruments to spur innovation.[7] They show that there is strong evidence that R&D tax credit policies are powerful and efficient tools to encourage private R&D. They also conclude that there is still a need for more evidence on the effect of university and more generally research funding on private sector outcomes. In one of the important contributions in this literature, Azoulay et al. (2019) link NIH grants with the publications they generate and the patents in the biotechnology and pharma industries that cite those publications. Using an identification strategy based on the NIH funding rules, they show that a $10 million boost in NIH funding leads to a net increase of 2.3 patents. With a spatial focus, Hausman (2021) studies how universities can be a driver of industrial agglomeration and shows that after the Bayh Dole Act, the industries closest technologically to the local university witness a growth in employment and innovative outcomes. Our contribution is first to propose a new measure of proximity. Rather than assigning a technological position to academic groups as in Hausman (2021), we assign a scientific position to firms. Second, we propose a different identification strategy, based on direct financing of academic clusters, to estimate the impact of public research funding on private sector innovation. The program we use can be seen as a middle ground between the project-specific funding used in Azoulay et al. (2019) and the university funding in Hausman (2021).[8] In contrast to Azoulay et al. (2019), the program we study applies to all fields, and not only the biotech and pharmaceutical industries which have been shown to be particularly sensitive to university spillovers (notably in their location choice, see Abramovsky, Harrison, and Simpson, 2007; Abramovsky and Simpson, 2011). Third, we provide evidence on channels through which these spillovers occur.

The literature on the local effects of academia was initiated by Jaffe (1989), which found strong effects of universities on corporate patenting, with some geographic dimension. Kantor and Whalley (2014) use national shocks on stock-return, affecting the value of university endowments, to instrument university spending, and found modest but significant local effects on non-research wages. Bikard and Marx (2020) study the importance of hubs in the use of academic science by firms. Akcigit, Hanley, and Serrano-Velarde (2021) find that basic research has broader spillovers than applied research and that subsidizing basic research achieves a bet-

7. As noted by Bloom, Van Reenen, and Williams (2019), other tools financing directly innovative firms have proved to be efficient. In particular, research grants have been shown to efficiently stimulate innovation among directly targeted firms (Howell, 2017) and to generate substantial spillovers both across geographical and technological spaces (Myers and Lanahan, 2022).

8. In fact the funding of such academic clusters, based on themes proposed by researchers themselves, is a growing instrument that appears promising. Our work leads us to introduce a new measure of proximity discussed in detail in Section 3.

ter allocation of research efforts.[9] There is also a literature studying the local impact of public spending in R&D, notably for military purposes, which typically relies on the comparison of areas more or less exposed to public procurement shocks efforts (Moretti, Steinwender, and Van Reenen, 2019; Kantor and Whalley, 2022).

An important contribution of our paper is also to provide empirical evidence on the channels through which spillovers occur, both exhaustive evidence from official reports by academic groups and causal evidence on certain channels of spillovers. There is a large literature that discusses the question, but it is mostly based on survey of companies.[10] There is evidence consistent with all the channels mentionned above: Cohen, Nelson, and Walsh (2002) highlight the importance of the informal channel (publication of papers, participation to conferences and interpersonal exchanges), De Fuentes and Dutrénit (2012) find that the most important channels are common R&D projects, property rights, and human resources sharing.[11] Rather than using surveys, we exploit official reports of all the financed units. Moreover, to our knowledge, this is the first paper to provide causal evidence on channels.

The paper is structured as follows. In Section 2, we present the LabEx funding program and our main data sources. In Section 3 we present and discuss our proximity measure, and provide a number of validation checks. Section 4 presents our main results on the impact of the funding program on exposed industries. Section 5 studies the channels through which spillovers occur. Section 6 assesses the policy implications of our results and concludes.

2 Data and identification

2.1 LabexEx program

Based on a bipartisan report written by two former prime ministers, the French president, Nicolas Sarkozy, announced in 2009 a large-scale investment plan for research and productivity, the *"Plan d'Investissement d'Avenir"*. One important component of this initiative was the LabEx program, aimed at financing consortia of research units that planned to work on a common theme

9. Arora, Belenzon, and Sheer (2021) focus on spillovers from corporate science to corporate inventions, and find that such spillovers must be particularly large, as firms are very sensitive to what might benefit their competitors in their investment decisions in science.

10. See for instance Perkmann et al. (2013) and Ankrah and Omar (2015) for reviews of this literature.

11. Agrawal and Henderson (2002) focus on transfers from MIT research, and find that only 10% of knowledge transfers passes through patents, making fundamental research outputs (such as the academic papers we use) very important.

(what we refer to as an academic cluster or LabEx).[12]

The program was run on a bottom-up and fully competitive basis at the national level. A first call for proposals was issued in 2010. Each application involved several research units with one coordinator in charge. The 241 applications received were sent to external reviewers and an independent international committee selected 100 winning proposals that were announced on March 25, 2011. In response to the second call for proposals made in October 2011, 71 were funded out of the 195 submissions (including 55 resubmissions from the first stage). The funding for these academic clusters was for an 8-year period (potentially renewable), with an average allocation of 10 million euros, ranging from 2 to 30 million euros, paid through yearly transfers. In 2019, the LabEx were evaluated by an international jury, which recommended that 11 not be renewed.

The stated goal of the program was to favor the emergence of ambitious scientific projects, to spur the production of academic papers and make these clusters visible on the international scene.[13] The labs were also encouraged to reach out to the local communities, including private firms. This was however a secondary goal, corresponding to one out of the seven criteria the jury had to evaluate.[14] As shown in Table A2, the corresponding grade did not have a significant impact on the probability of being selected, in line with the idea that the international jury, composed of academics with limited knowledge of the local French industrial sector, had a harder time evaluating this criterion.

LabEx data The ANR (*"Agence Nationale de la Recherche"*, the institution that supervised the LabEx program) shared with us the application files they received.[15] All files include the name of the coordinator, the name and identifying codes of the partner research units, the amounts requested, the funding decision and a summary of the project.[16] In addition each file contains a bibliography that we use to build our exposure measure. Finally, The ANR provided us with the grades for each proposal, including the rejected projects, information we use to conduct robustness exercises.

12. Similar policies have been developed in other countries such as Germany and Nordic countries, with a similar goal of supporting and developing a limited number of world class research clusters.

13. Carayol, Henry, and Lanoë (2020) study the effect of the policy on research output and show that it increased co-publications between members of the funded LabEx.

14. The criterion was "Potential of the research project in terms of innovation and impact". The other six criteria are the quality of the teams and facilities, the relevance of the research project goals, involvement in academic training, organization and management, institutional strategy (universities and research institutes), and project/means adequacy and ability to generate resources.

15. The ANR shared with us 200 of the 241 files for the 2010 call, removing the proposals that received the lowest grade. For the 2011 call, they shared all the files with us.

16. For confidentiality concerns, we were not given access to the full text of the proposal.

2.2 Data sources

We have assembled a large variety of data sources, allowing us to explore spillovers and the mechanisms underlying them. The data are used to (i) construct a measure of scientific proximity (data on private patenting and publications listed in the bibliography of proposals), (ii) provide causal evidence on spillovers (data on the LabEx program), (iii) show evidence on mechanisms (administrative employer–employee data to track labor mobility, data on co-supervision of PhDs and data on subcontracting) and (iv) to compare the LabEx program with other instruments to spur private sector innovation (data on the French tax credit program CIR). To guide the reader, we provide in Table 1 the list of variables and the source used to construct them. More details on the data sources are given in Online Appendix C.[17]

Patent data We rely on Patstat (Spring 2020 Edition), a database produced by the European Patent Office which contains all the patent applications filed in most intellectual property offices in the world. Since these applications are entered into the database with some lag, the 2020 edition provides exhaustive coverage of filings up to 2015. We match French companies with their patents in the database. The matching procedure is described in Appendix C.1. In addition, we use the city of the inventors in the OECD REGPAT database (Maraut et al., 2008), July 2021 edition, to geocode the patent and allocate it to a commuting zone. Finally, we augment this database with information on citations to the non-patent (mostly academic) literature. This is done using PatCit (Cristelli et al., 2020), an open-source database aiming at retrieving all citations made within patent applications, including those that only appear in the text (details provided in Appendix C.1). The patent data is used to measure the technological distance between firms and academic clusters, as well as an outcome variable in our analysis.

Linked employer-employee data (DADS) The DADS Postes is an administrative dataset which contains, for each employment spell in France, both for the current and preceding year, the identity of the employer, the wage, the hours worked, the type of occupation, the city of work. We use this information to construct our key measure of spending on R&D based on the total wage bill of engineers.[18] Since we know the city of work, this measure can be defined very precisely at the local level. We show in Appendix C.2, using alternative sources, that this variable is a good measure of overall R&D employment.

17. Note that access to confidential data, on which this work is based, has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD).

18. Identified through positions with an occupation and socio-professional category (PCS) beginning with 38: "Engineers and technical managers of companies".

The DADS is also used to measure mobility from the public research sector to the private sector. We exploit the fact that we know the occupation in year $t-1$ of the worker.[19] We distinguish movements by junior researchers (those in PhD in the preceding years) versus mobility of more senior researchers.

Research tax credit data (GECIR and MVC CIR) France has a large R&D tax credit program called CIR (institutional details provided in Online Appendix C.3). We obtained from the tax authority DGFiP and the higher education ministry MESRI, the datasets called GECIR and MVC CIR, which record the filings made by firms on their R&D expenditures, used to determine the fiscal transfers. The research tax credit is declared at the company level by the fiscal parent company, and since our geographical unit of observation is the commuting zone, we need to allocate the total amount claimed. We do so according to the share of the company's engineers in the commuting zone (the procedure is described in Appendix C.3). This dataset is used to measure outcomes such as total spending on R&D but also to identify specific channels, exploiting the information on outsourcing to public research labs that firms need to report when they claim tax credits.

PhD cosupervision data (Cifre) France has a public subsidy program, called Cifre, for co-supervisions of PhDs between a public lab and a company. The two parties sign a contract that specifies how the student will share her time between the two institutions and what will be the rules regarding intellectual property (see Appendix C.4 for details on the institution). We obtained data on all Cifre contracts at the individual level, where we can identify the collaborating firm with the national firm identifier, the municipality where the PhD student is employed and the public research lab co-supervising the student. These data are available from 2003 to 2018. This measure of PhD co-supervision is used in the section on channels.

Academic spinoffs (JEU) France has in place a program of payroll and tax exemptions for academic spin-offs launched by students or faculty members in universities (JEU, *Jeunes entreprises universitaires*). We obtained data on firms involved in this programs as described in Appendix C.5.

Plant register (REE) We use yearly information on the stock of firms and establishments from the French statistical office (Insee). Using this source, we can calculate the number of new plants

19. The public sector was included in the DADS from 2009.

opened each year. These plants can be created either by existing firms or by new entrants.

Table 1: Description of variables

Variable name	Source	Coverage	Details
Variables used for main results			
R&D wage bill	DADS	2005-2018	Sum of wages of employees in PCS 38
R&D hours	DADS	2005-2018	Sum of hours worked by employees in PCS 38
R&D hourly wage	DADS	2005-2018	Hourly wage of employees in PCS 38
Total R&D claims	GECIR	2008-2018	Sum of R&D claims declared in the CIR (R&D tax credit) program
Number of patents	PATSTAT	2005-2018	Number of patents
Number of new plants	REE	2005-2018	Number of new plants
Variables used to study channels			
PhD co-supervision	Cifre	2005-2018	Number of Cifre (PhD co-supervisions)
Academic spin-offs	JEU	2009-2018	Number of young academic private ventures entitled to tax breaks
Outsourcing R&D to public labs	GECIR	2008-2018	Amount of outsourcing to public labs declared to claim CIR
Transfer of senior academics	DADS	2010-2018	Number of transfers of senior academics from a main job in academia to a main job in private sector
Transfer of junior academics	DADS	2010-2018	Same for junior academics
Transfer of researchers	DADS	2010-2018	Same for total number of transfers
Hiring of young PhDs	GECIR	2008-2018	Number of recent PhDs hired as declared in the CIR tax declarations
Collaboration with univ.	CIS	2004-2016	Average probability to collaborate with a local university.

3 Measurement of proximity and identification

One of the contributions of this paper is to propose a novel measure of the proximity between a given industry and a research group. The key idea is to measure whether the science that an industry uses and the science that a public lab produces coincide, without restricting to existing direct links. This procedure thus infers a scientific position of each industry. It is in contrast with the more common in the literature which attributes a technological position to each university, based for instance on the subset of knowledge it produces as academic patents.[20] In this section, we first present the measure before discussing its properties and the relation with the literature.

3.1 Construction of measures of proximity and exposure

To capture the idea of proximity in the science produced by an academic group and the science used by a sector, we exploit publications in academic journals. Define s_{lj} as the share of papers from academic group l published in journal j and s_{ji} the share of citations to journal j made by industry i. We define our measure of proximity as;

$$\text{prox}_{li} = \sum_j s_{lj} \cdot s_{ji} \tag{1}$$

i.e. the sums the product of the shares over journals j.[21]

Based on this measure of proximity, we construct the measure of exposure of industry i in commuting zone k to a funding shock, expo_{ik}. The measure is the sum over all LabEx in the commuting zone k of the funding received by each LabEx, weighted by the proximity of the LabEx to the industry. We have:

$$\text{expo}_{ik} = \sum_{l \in k} d_l \cdot \text{prox}_{li} \tag{2}$$

where d_l is the amount of funding received by the LabEx proposal l. [22]

20. In a subsequent work (Bergeaud et al., 2022), we show that this novel measure performs better in detecting local spillovers from universities than measures based on academic patents or direct citations.

21. Although our measure does not require direct links between a group and a firm, the intuition for our measure can be presented in the following way. The paper in journal j may trigger an idea of a technological application, which will originate from industry i with a probability equal to the share s_{ji} of citations to journal j made by industry i. The proximity measure thus represents the probability that the scientific production of the lab l is used by firm i.

22. expo_{ik} can be interpreted as the amount of funds implicitly directed toward firms of industry i in commuting

In our particular setting, in order to compute the proximity and exposure measures between LabEx and industries, we use the unique information contained in the bibliographies of the proposals. We tagged bibliographic references[23] in these applications, and linked them to the journals in which they were published. We then characterize each LabEx l by the vector of shares s_{lj} of papers in the bibliography published in each journal (ISSN). For instance, a LabEx project on aerospace engineering could be characterized by half of the bibliography being published in *Progress in Aerospace science*, the other half in *Journal of Fluid Mechanics* and zero in all other journals, while a LabEx project in molecular biology might have a vector composed of a third of its publications in *Journal of Biological Chemistry*, a third in *Cell*, a third in *Journal of Molecular Biology*, and zero in all other journals. These vectors of shares will therefore finely represent how a project is positioned in the scientific space.[24] On the firm side, we take the universe of patents owned by French firms before 2011 and use the available DOIs in the PatCit database to determine which academic articles are cited by these patents. We then link these articles to the journal in which they were published. This allows us to compute the share s_{ji} of citations to journal j made by industry i.

3.2 Proximity measure: discussion

3.2.1 Validation

Given the novel nature of this indicator, we start by providing some evidence on its validity in Online Appendix D, before discussing the relation with other indicators in the literature. We first exploit the initial reports we obtained for the funded projects, which sometimes mention potential collaborations. For each LabEx, we can thus determine the sectors that are mentioned in the reports. We show that sectors that appear more in the reports are closer, according to our proximity measure, to the LabEx. In a second exercise, we use the Community Innovation Survey (CIS) which surveys more than 10,000 companies every two years on the nature of their

zone k, as part of the Labex program. It can be compared to other innovation subsidies received by firms, for instance through the research tax credit, an exercise which we conduct in Section 6.

23. We used the machine-learning library Grobid.

24. Most journals are highly specialized in a given scientific field but a small number are more generalist or interdisciplinary. Table A4 in Appendix A shows that our results are not impacted when we remove such academic reviews. Formally, we use Crossref to assign a scientific category to each article (either using a broad classification into 18 fields, or a more detailed one using 210 fields). We then calculate a Herfindahl index of concentration to select generalist journals. Alternatively, we also use the detailed classification to aggregate journals into scientific categories from which we construct weights s_{li} and s_{ji}. Here again, our main result holds but is less precise, suggesting that using the full variety of academic journals provides a more detailed measure of proximity between sectors and LabEx.

innovative activities. As in Abramovsky, Harrison, and Simpson (2007), we use the question on the importance of sourcing from universities and higher education institutions by the firm. We show that sectors in which a high share of firms report using the research published by the public sector prior to 2011 are those with a higher average proximity. Both these validation exercises are presented in Appendix D.

3.2.2 Links with other measures

Our measure of proximity captures the distance between the science produced in academia and the science used in the industry to develop technological applications. It thus infers a scientific position for each industry, which is different from the more standard approach used in the literature which attributes a technological position to each university using patents (and their distribution across technological classes). A first stream of papers (for instance Hausman, 2021) directly uses academic patents (i.e. patents filed by public labs) to infer the technological position of universities. A second approach is to use the patents which directly cite the papers produced by a research group, as in Azoulay et al. (2019) in the context of the biotechnology and pharmaceutical industries.

We believe these two approaches typically used in the literature are not ideal in our context. First, academic patenting is a rather rare event in France prior to 2010.[25] More importantly, we want to exploit a shock in funding that could affect the production of the academic group and the way the knowledge produced is used in the industry. Thus, using a method that relies on citations to academic papers prior to the shock could be problematic. For instance, the LabEx members, prior to funding, might not have invested time in collaborating with the industry. They might have produced science useful for firms, but knowledge not yet exploited by the industry, therefore having received very few citations.[26] On the contrary, unless the funding changes the type of science produced by the LabEx, the measure of proximity we propose should not be affected by the funding.

We nevertheless compare our approach to alternative measures based on patents. The detailed construction is presented in Section 4.2. We identify patents which directly cite papers contained in the bibliography of a given LabEx project and compare these to the set of patents filed by an industry. These alternative measures of proximity and our proposed measure are

25. Moreover, university patents have been shown to capture only a small share of the effects produced by universities' knowledge production (Agrawal and Henderson, 2002), and may therefore only reflect knowledge transfers of a very specific kind.

26. Or simply the group might have produced lower quality work not useful for the industry, before they received the funding.

positively but not perfectly correlated, suggesting that they capture different notions of scientific proximity. Moreover, we show in Section 4.2 that our main results are less precise but still hold when we replace our baseline exposure with exposure based on these alternative measures of proximity.

3.3 Identification

We exploit the shock in funding resulting from the LabEx program that affected certain commuting zones and not others after 2010 (the start of the program). Within these commuting zones, certain industries were exposed to the shock because of their technological proximity to the funded academic cluster. The variations in relative exposure across space and industries allow us to control for shocks specific to the commuting zone. Therefore, our identifying variation comes from differential exposure to the policy of industries within a commuting zone.

Specifically, for a given industry i in a commuting zone k in year t, we are interested in an outcome variable Y_{ikt}, such as employment in R&D. We estimate the following model:

$$Y_{ikt} = \beta \times \mathbb{1}\{t > 2010\} \times \ln(1 + \text{expo}_{ik}) + \alpha_{ik} + \delta_{tk} + \varepsilon_{ikt} \tag{3}$$

where expo_{ik} measures the exposure of industry i in commuting zone k to the funding shock, as introduced above, a measure which is time–independent. The parameter α_{ik} is an industry $\times$ commuting zone fixed effect, while $\delta_{t,k}$ captures flexibly commuting zone specific time trends. The parameter of interest is β and captures the impact of exposure on outcome variables. The underlying assumption is that the more exposed industry–commuting zone dyads would have followed similar trends as the less exposed ones, absent the funding shock. To increase the likelihood that this assumption is satisfied, we restrict our sample to commuting zones that had at least one LabEx proposal submitted in the competition and industries with non-zero proximity to the local research cluster proposal in at least one commuting zone. Commuting zones where no proposal is submitted are typically much more rural and less active in research than those in our sample (see Figures E2 in Online Appendix E).

We also present the results graphically by estimating a dynamic difference-in-differences specification allowing us to gauge the magnitude of effects over time:

$$Y_{ikt} = \sum_{\substack{d=2005 \\ d \neq 2010}}^{2017} \beta_d \times \mathbb{1}\{t = d\} \times \ln(1 + \text{expo}_{ik}) + \alpha_{ik} + \delta_{t,k} + \varepsilon_{ikt} \tag{4}$$

The estimated coefficients β_d can be causally interpreted under the identifying condition

that the treatment is orthogonal to the error term in equation (4) conditional on CZ $\times$ sector and CZ $\times$ year fixed-effects. Formally, this identifying assumption writes as:

$$\mathbb{E}[\varepsilon_{ikt}(\mathbb{1}\{t = d\} \times \ln(\text{expo}_{ik} + 1)) \mid \alpha_{ik}, \delta_{t,k}] = 0 \quad \forall (t, d)$$

The identifying assumption states that, in the absence of the policy reform, the outcome variable would have been similar, within a given commuting zone, among industries more or less exposed to spillovers from the reform. This common trend assumption cannot be directly assessed. However, finding β_t not to be significantly different from zero for $t < 2010$ is evidence consistent with the absence of differential pre-trends between differentially exposed industry-commuting zone dyads.

The identification strategy rests on two sources of variation: variation in clusters that were selected and variation in industries exposed to the funding shock. Both sources of variation give rise to specific concerns, that we address with tests presented in Section 4.4.

First, even in the presence of exogenous shocks, exposure of industries to shocks might not be random, and the unobservable variables explaining exposure might also drive the dynamic evolution of these industries. To address this concern, we follow Borusyak and Hull (2021) and consider counterfactual realizations of the shocks in a number of robustness exercises. We use the data on grades and amounts requested for all submitted projects to compute the funding non selected clusters would be expected to obtain, had they been selected.

The second threat might come from the selection process of the funded units. One might worry that the selection was performed based on how connected the proposed clusters were to potentially booming sectors in a specific geographical area. While we view this event as unlikely given that the jury was made of international experts with no specific knowledge of the French economy and was asked to judge purely scientific quality, we however provide further tests. In particular, using the key information on grades, we restrict the sample to show that this mechanism is not at play.

4 Spillovers from public to private

4.1 Main results

The private sector can benefit from research in the public sector if it manages to integrate the innovations and ideas produced by public researchers in its production process. It can also benefit by building on these ideas to produce its own innovations and new products. Both

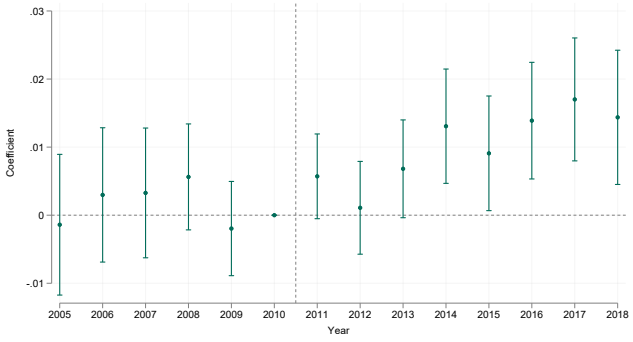
processes require additional spending on R&D inputs, in the first case to increase absorptive capacity and in the second to be in a position to innovate. In this section, we estimate specifications (3) and (4) for different dependent variables, to explore the causal effect of the shock in funding on R&D inputs and outputs.

4.1.1 R&D efforts

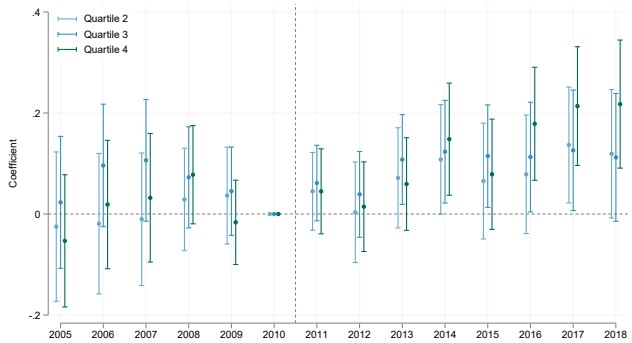
We first study how the financing of public research affects private R&D efforts. Figure 1 presents the results of the estimation of the dynamic model (equation (4)) using as outcome variable the total R&D wage bill (in log).[27]

Figure 1: Impact of Labex funding on employment in R&D

(a) Average



(b) By quartile of exposure



Notes: Panel (a) presents point estimates and 95% confidence intervals of coefficients β_d from equation (4) for values of d ranging from 2005 to 2018. β_d has been standardized to 0 for $d = 2010$. Estimates were obtained through OLS with standard errors clustered at the industry–commuting zone level. The dependent variable is the log of total engineer wages in plants of a given commuting zone–industry pair. Panel (b) presents a similar figure but the coefficient β_d is replaced by a binary variable for different quantiles of the level of exposure, where the reference category is the set of industry–CZ pairs below the first quartile of non-zero exposures. 42301 obs (3761 industry–CZ pairs).

We present the results using the continuous measure of exposure in Figure 1(a) and we show the effect by quartile of exposure in Figure 1(b). The results show that a few years after the treatment, industry–location pairs that were more exposed to the local public research funding shock witnessed an increase in spending on R&D labor. 5 years after treatment, we observe an increase of around 1.5% in R&D wage bill when the exposure of an industry in an commuting zone is doubled. This Figure also shows the absence of differential pre-trends in the years building up to the financing, which is evidence in support of our identification strategy. Figure 1(b) shows that the effect is mostly driven by pairs of industry-location that are in the top

27. As explained in the previous section, we use the total R&D wage bill as a proxy for R&D expenditures, since this is a quantity that is precisely observed at the local level for the full set of firms (see discussion in Appendix C.2).

quartile of exposure. On average, these pairs increase their spending on employment in R&D by 20% compared to the least exposed units (which received no funding).

The first line of Table 2 presents the corresponding static coefficient β from equation (3). We then decompose this effect in the next two lines and show that the increase in wages is explained for three fourth by an increase in hours worked, and for one fourth by an increase in hourly wage. In these three regressions, the static coefficient is significant and positive. The table also systematically presents the average value of pre-trend coefficients (β_d with $d < 2010$) from an estimation of equation (4), to test the absence of a trend prior to the treatment.

As an alternative measure of R&D effort, we use as our outcome variable the total R&D claims in the tax credit data (see Appendix C.3 for details). The results are presented in the fourth line of Table 2 and are similar to those found for total R&D worker wage bill. This is not surprising as the correlation between the two variables is very high (see Table C1 in Appendix C.2), even though the coefficient on total R&D is larger.

Table 2: Main results

	Static Coefficient	Obs.	Pre Trends
R&D wage bill (log)	0.0087***	47,986 obs (4285 pairs)	0.0017
	(0.0033)		(0.0038)
R&D hours (log)	0.0068**	47,985 obs (4285 pairs)	0.0030
	(0.0031)		(0.0036)
R&D hourly wage (log)	0.0019**	47,985 obs (4285 pairs)	-0.0013
	(0.0007)		(0.0009)
Total R&D claims (log)	0.0150**	27,373 obs (3073 pairs)	0.0022
	(0.0064)		(0.0076)

Notes: Each line corresponds to a different dependent variable. Column 1 shows the coefficient from a static difference-in-differences specification run over the period 2005–2018 (see model (3)). The reported coefficient corresponds to the exposure variable interacted with a post (i.e. after 2010) dummy variable. The last column shows the average value of the pre-trend coefficients of the model (4), estimated over the same period 2005-2018. The unit of observation is a pair of commuting zone × 5-digits industry. All models include a commuting zone-industry fixed effects and a set of commuting zone-year dummies. All estimations use the OLS. Standard errors are clustered at the pair level. The number of observations is lower for Total R&D claims because a larger number pairs of CZ and sectors report 0 R&D.

Discussion on magnitude and external validity. Results presented in Table 2 imply quantitatively important effects. They suggest that a one standard deviation increase in exposure (≈ 5.05) translates into a 3.4% increase in R&D (engineer) hours worked. Alternatively, we can compute the euro for euro impact of the program on R&D wagebill. We first use the estimate (0.0087), the average value of the exposure (5.24) and the average pre-reform R&D wagebill (5.6 million euros) to obtain a predicted average effect on the wagebill in euros and divide it by the average cost of the program. We end up with a ratio of 0.72, meaning that one euro of

financing to the LabEx program generated on average a 72 cents increase in the R&D wagebill in the medium-run. A similar calculation gives an increase of private R&D expenditures of 98 cents for one euro of funding.

These estimates are obtained using cross-sectional variation in exposure across industries within CZ. Therefore, they cannot be directly extrapolated to assess the *aggregate* effect of the policy on R&D employment. In particular, it is possible that the policy reallocated engineers from low to high-exposure industries within commuting zones. We try to assess the magnitude of such displacement effect in Table A3 of the online appendix. In this table, we decompose R&D hours worked at time t depending on where workers were employed at $t-1$. We distinguish in particular hours worked by incumbents, defined as workers employed in the same CZ $\times$ industry at $t-1$, movers from different industry and/or CZ and new entrants on the labor market. We see that the bulk of the overall effect is driven by incumbents—which could reflect a higher retention rate, potentially driven by the positive wage effect we estimate—with no effect on movers coming from different industries within the same CZ—which corresponds to the identifying variation. Overall, while we cannot completely rule out that our diff-in-diff estimates reflect in part displacement from the control group, this exercise suggests it is not a first-order component of the overall effect.

4.1.2 R&D outputs

Does this increase in spending on R&D inputs translate into outputs, such as the creation of new plants and production of new patents?

To measure the effect of the LabEx program on the creation of new plants, we use the registry of plants administrated by the Insee and calculate the number of new establishments in each CZ and each sector. Table 3 shows a positive effect of LabEx exposure on the probability of creation of a new plant in an industry–CZ pair, corresponding to an increase of 0.2 percentage points when exposure doubles. Decomposing this effect between plant creation from new and incumbent firms shows that it stems almost entirely from new firms.

The impact on patenting is explored in the second part of Table 3. As is standard in the literature, we use a Poisson regression to estimate the coefficient, which takes into account the very large number of observations with no patent. The baseline model yields an insignificant static coefficient along with non zero pre-trend. However, when sector-year fixed effects are added to the model, the coefficient becomes positive and significant and the pre-trends are no longer significant. This suggests that industry-specific patenting dynamics play a first-order

Table 3: Additional outcome results

	Static Coefficient	Obs.	Pre Trends
Creation of new plants (binary)	0.0021***	59,990 obs (4285 pairs)	0.0006
	(0.0008)		(0.0014)
Creation by new firms	0.0019***	59,990 obs (4285 pairs)	-0.0017
	(0.0007)		(0.0014)
Creation by existing firms	0.0008	59,990 obs (4285 pairs)	0.0021
	(0.0008)		(0.0015)
Number of patents	0.0148	17,456 obs (1248 pairs)	-0.0212**
	(0.0245)		(0.0108)
Number of patents (with sector FE)	0.0443**	15,657 obs (1232 pairs)	-0.0265
	(0.0184)		(0.0219)

Notes: Each line corresponds to a different dependent variable. Column 1 shows the coefficient from a static difference-in-differences specification run over the period 2005–2018 (see model (3)). The reported coefficient corresponds to the exposure variable interacted with a post (i.e. after 2010) dummy variable. The last column shows the average value of the pre-trend coefficients of the model (4), estimated over the same period 2005-2018. The unit of observation is a pair of commuting zone $\times$ 5-digits industry. All models include a commuting zone-industry fixed effects and a set of commuting zone-year dummies (the last model also include a set of 5 digit industry-year fixed effects). Estimations on entry use an OLS estimator while estimations on patents uses a Poisson model. Standard errors are clustered at the pair level.

role over our period of study and confound our baseline estimation.[28]

Overall, these results show that investments in public research have a causal and significant medium term effect on R&D spending and R&D outputs in local industries that are scientifically connected to the research entity. We explore the robustness of these results in the next section.

4.2 Robustness

4.2.1 Additional controls and alternative samples

In our main specification, we include flexible commuting zone–time effects, but do not include industry–time effects. Our first robustness exercise is therefore to add these flexible time effects specific to each industry (2-digits, 88 categories). The results, presented in Figure B2, are very similar to those in Figure 1, suggesting that spillovers across commuting zones are not that important, but the standard errors increase. We replicate Table 2 adding these fixed effects and show that the results are overall preserved (see Table 4). Going one step further, we also add 5-digit industry-year fixed effects to the model. The resulting coefficient is shown in Table 4 for the main dependent variable (the total wage bill of R&D workers taken in log). Its magnitude is very similar to the one of the baseline estimation and its precision is slightly lower, but still

28. It is well known that the propensity to patent is very different across technologies and therefore across sectors. If this propensity has evolved over time, then adding industry-year fixed dummies will capture this effect.

significantly different from 0.

A possible concern is that LabEx are widely concentrated in the Paris area (*Ile de France* region) and in specific sectors (chemistry and pharmaceutical in particular, see Figure 3). In addition, the list of 2 digit industry codes includes an "R&D sector" which also accounts for a large share of the total exposure. We check that our results are robust to removing these specific observations. We thus alternatively restrict the sample by removing the R&D sector, then the chemistry and pharmaceutical industries and finally the Paris region. Our main coefficient of interest is barely affected (see Table 4).

Table 4: Robustness checks

	Static Coefficient	Obs.	Pre Trends
Baseline	0.0087***	47,986 obs (4285 pairs)	0.0017
	(0.0033)		(0.0038)
Adding sector fixed-effects			
1. 2-digit sector FE	0.0107***	47,986 obs (4285 pairs)	-0.0018
	(0.0038)		(0.0044)
2. 5-digit sector FE	0.0097*	47,986 obs (4284 pairs)	0.0029
	(0.0041)		(0.0048)
Removing highly exposed sectors/locations			
3. Remove R&D sector	0.0083**	47,157 obs (4214 pairs)	0.0018
	(0.0033)		(0.0038)
4. Remove pharma and chemical sectors	0.0102***	44,987 obs (4009 pairs)	0.0027
	(0.0034)		(0.0040)
5. Remove Paris region	0.0084**	39,400 obs (3544 pairs)	0.0037
	(0.0035)		(0.0040)
Alternative measures of proximity			
6. IPC3 weights	0.0068*	47,986 obs (4285 pairs)	-0.0013
	(0.0037)		(0.0042)
7. IPC4 weights	0.0060*	47,986 obs (4285 pairs)	0.0001
	(0.0036)		(0.0040)
8. Embedding weights	0.0061*	47,986 obs (4285 pairs)	-0.0007
	(0.0031)		(0.0036)

Notes: This Table presents the results of the same estimation as in Table 2, using as dependent variable the log of the total wage bill of engineers, and either adding sector specific trends in lines 1 and 2, or applying restrictions to the data in lines 3-5 (see Section 4.2.1), or using alternative distance measures in lines 6-8 (see Section 4.2.2).

4.2.2 Alternative measures of proximity

As explained in Section 3.2, the typical measure of proximity used in the literature would assign a technological position to the LabEx and compare it with the position of different sectors. To construct this alternative measure of proximity, we proceed as follows. First, we use the PatCit database to identify the patents which directly cite papers contained in the bibliography of a

given LabEx project. This procedure assigns a portfolio of patents to each LabEx that can be compared to the set of patents obtained by each industry. We then compute two different types of proximity metrics. The first one is based on the technological classes (IPC) of the patents. For each LabEx and for each industry, we calculate a vector of weights on each IPC class at the 3-digits level (there are 123 such technological classes in our data) and simply take the Euclidean distance between each pairs of LabEx-industry. We repeat the procedure at the 4-digits IPC level. The second type of metric uses the embedding of each patent as calculated by Google (Srebrovic, 2019).[29]

These three alternative measures of proximity are correlated with our baseline (the correlation are respectively of 0.44, 0.59 and 0.60 with the Embedding, IPC3 and IPC4 measures) and we use them to construct three different exposures. We then run our main specification using the three new measures. Results shown in Table 4, lines 6-8, and are qualitatively similar to those of our baseline model with slightly smaller and less precise coefficients (see Figure B3 in the Appendix).

4.3 Localization of spillovers

Our proximity measure is based on the premise that spillovers are local and occur withing the boundaries of the commuting zones. This assumption is motivated by the local dimension of spillovers that is highlighted in the literature. We put this assumption under scrutiny in Online Appendix E. First, we augment specification (3) to include the exposure of neighboring CZ as a control variable. We find that this additional variable does not explain variations in total R&D wage bill. Second, we define an alternative exposure measure where all locations are potentially affected by funding shocks, the effect decaying with distance. We use different parameters that govern the strength of this decay. All the estimates imply an increase in R&D spending after the treatment. This analysis can be found in Appendix E.

29. Embeddings are a learned representation of a complex object composed of many features with the goal of reducing its dimensionality. In this case, each patent has been associated with a vector of 64 real numbers computed using a machine learning model that predicted a patent's technological classes from its text. In other words, the embedding vector encodes the semantic content of a patent into an algebraic object from which we can easily calculate a distance. We calculate the unweighted average of the embedding vectors of each patents associated with a given Labex on the one hand and for each patent associated with a given industry on the other hand, and calculate the cosine distance between the two.

4.4 Placebo tests

As discussed in Section 3.3, our identification strategy relies on a parallel trends assumption for the pairs of sector and CZ differentially exposed to the shock. These pairs vary in exposure to funding along two dimensions: academic clusters within the different CZ obtain different levels of funding and within CZ different sectors vary in exposure to the treated units. Both these sources of variation may give rise to threats to identification, that we consider in turn.

4.4.1 Variation in the exposure of industries

To address the concern of possible non–random exposure of pairs of sector and CZ to the shock, we first follow Borusyak and Hull (2021), who have developed a new methodology that constructs counterfactual shocks by simulating the data generating process that assigns the funding to candidate LabEx.

Before applying Borusyak and Hull (2021), we first present an extreme exercise that illustrates their approach. We compute a counterfactual exposure as if rejected LabEx were in fact accepted, while the accepted ones were rejected.[30] We then estimate equation (3) and present the results in line 1 of Table 5. The results show that this counterfactual exposure measure does not have predictive power.

The method in Borusyak and Hull (2021) is based on estimating the effect of a large number of these counterfactual exercises. We apply their approach to our data. We have information on 268 candidate projects, 139 of them have been accepted and 129 have been rejected. We randomly draw winners across all candidates, keeping the share of accepted projects fixed. For each LabEx that have been allocated in the new treatment group, we then assign the average funding value observed for actually accepted LabEx $\bar{d}_l$ while for the other LabEx, we set the funding to 0. We can then construct the corresponding counterfactual measure of exposure at the sector-CZ level which results from this specific first permutation that we note $\widehat{\text{expo}}^{(1)}$.

We then replicate this procedure a thousand time to generate $\widehat{\text{expo}}^{(2)}$, $\widehat{\text{expo}}^{(3)}$,... $\widehat{\text{expo}}^{(1000)}$. From which we construct a control variable:

$$\widehat{\text{expo}}^{permut} = \frac{1}{1000} \sum_{p=1}^{1000} \widehat{\text{expo}}^{(p)}$$

30. Formally, we use our predicted measure of exposure (see below) for the group of rejected academic clusters and set it to 0 for the actual funded LabEx.

which we add as a control variable in equations (3) and (4) as shown in equations (5) and (6) respectively.

$$Y_{ikt} = \mathbb{1}\{t > 2010\} \times \left(\beta \ln\left(1 + \text{expo}_{ik}\right) + \gamma \ln\left(1 + \widehat{\text{expo}}_{ik}^{permut}\right)\right) + \alpha_{ik} + \delta_{tk} + \varepsilon_{ikt} \quad (5)$$

and

$$Y_{ikt} = \sum_{\substack{d=2005 \\ d \neq 2010}}^{2017} \left(\mathbb{1}\{t = d\} \times \left(\beta_d \ln\left(1 + \text{expo}_{ik}\right) + \gamma_d \ln\left(1 + \widehat{\text{expo}}_{ik}^{permut}\right)\right)\right) + \alpha_{ik} + \delta_{t,k} + \varepsilon_{ikt} \quad (6)$$

The results are presented in Table 5 (row 2). If the positive result reported previously is indeed a response to the funding, we would expect coefficients γ and γ_d to remain indistinguishable from zero, while coefficients β and β_d should be similar to those in the baseline model. The result shows that the actual exposure indeed continues to be positively associated with R&D effort after the treatment, while this is not the case for the counterfactual exposure.

We then exploit the richness of the data we obtained on the LabEx. For each proposal, including rejected ones, we observe the amount of funding requested, the grades obtained and the scientific field. We can therefore predict the funding that a LabEx would have received had it been accepted in the program: we estimate the following model for all accepted projects l:

$$d_l = \exp\left(\log(R_l) + \mu_{f(l)} + \nu_{n(l)} + t_l + \varepsilon_l\right)$$

where d_l is the funding actually received and R_l the amount requested. $\mu_{f(l)}$ is a vector of dummy variables for each scientific field and $\nu_{n(l)}$ a dummy vector for each grade category. Finally, t_l is a binary variable that takes the value 1 if the application has been filed in 2011 (as opposed to 2010). The coefficients are estimated using a Poisson estimator and used to predict $\hat{d}$ for all projects, including those that have been rejected. We then apply the exercise suggested in Borusyak and Hull (2021) but use the project–specific predicted value of the funding $\hat{d}_l$ to construct the counterfactual shock. The results, presented in Table 5 (line 3) are very close to those obtained in the first exercise.[31]

31. Borusyak and Hull (2021)'s approach also allows to conduct robust randomization inference based on the distribution of the coefficients obtained across counterfactual shocks permutations in order to test $\beta = 0$. The two exercises implies a p-value, defined as the probability that a simulated coefficient is larger than the baseline one when equation (4) is estimated using $\widehat{\text{expo}}^{(k)}$ instead of expo, of respectively 0.009 and 0.011. This approach to inference presents the advantage of accounting for the potential dependence across observations i, k induced by

Finally, to control for the fact that all candidate LabEx do not have the same likelihood of being funded, we replicate the exercise presented in line 2 of Table 5 but randomize the assignment within 4 categories of grades (below 27, between 27 and 29, between 30 and 32 and above 32, out of 35). Hence, we build a counterfactual exposure that keeps the proportion of funded LabEx within each category of grades unchanged. The results are presented in line 4 of Table 5 and again show a non-significant counterfactual coefficient while the actual exposure continues to be positively associated with the post treatment wagebill of engineers.

Table 5: Placebos, selection on grades

	Static Coefficient	Obs.	Pre Trends
Exposure of industries			
1. Counterfactual Exposure (rejected proposals)	0.0032	47,986 obs (4285 pairs)	0.0035
	(0.0035)		(0.0039)
2. Counterfactual Exposure (average)	-0.0050	47,986 obs (4285 pairs)	0.0035
	(0.0070)		(0.0079)
Actual Exposure	0.0125**		-0.0010
	(0.0057)		(0.0065)
3. Counterfactual Exposure (predicted)	-0.0054	47,986 obs (4285 pairs)	0.0039
	(0.0058)		(0.0080)
Actual Exposure	0.0127***		-0.0012
	(0.0058)		(0.0065)
4. Counterfactual Exposure (clustered)	-0.0003	47,986 obs (4285 pairs)	0.0024
	(0.0181)		(0.0203)
Actual Exposure	0.0087**		0.0015
	(0.0036)		(0.0042)
Selection of clusters			
5. Actual Exposure (average grades)	0.0063*	39,131 obs (3439 pairs)	0.0026
	(0.0035)		(0.0040)
6. Actual Exposure (outstanding grades)	0.0114***	36,081 obs (3185 pairs)	0.0019
	(0.0036)		(0.0042)

Notes: This Table shows the coefficients and standard errors of various estimations. The dependent variable is the logarithm of the total wage bill of engineers in each pairs of CZ and industry. Lines 2 to 4 correspond to the estimations of equations (5) and (6) using different measures of the counterfactual exposure as explained in Section 4.4. Lines 1, 5 and 6 show the results of the estimation of equations (3) and (4). Line 1 uses a measure of the exposure based on the predicted funding of the rejected LabEx. Lines 5 and 6 select on project with average grades (Grades between 26 and 32 for proposals filed in 2010 and between 30 and 32 (out of 35) for proposals submitted in 2011, there are 115 such proposals) and outstanding grades (greater than 32 out of 35, there are 73 such proposals). The specifications are otherwise similar to the one presented in Table 2.

the fact that variation in exposure derives from random funding decisions at the Label (l) level.

4.4.2 Variations in the selection of clusters

As discussed in Section 3.3, the second potential challenge to identification relates to the selection process. One might worry that the choice of one proposal over another was based on the motivation that the chosen LabEx was more connected to potentially booming sectors in a specific geographical area. While we view this event as unlikely given that the jury was made of international experts with no specific knowledge of the French economy, we however use the information on grades and restrict the sample to show that this mechanism is not at play.

First, we restrict attention to proposals that had a "standard" or "average" grade, i.e. the common grade support across accepted and rejected proposals.[32] This allows us to remove proposals that were either not good enough so that they had no chance to get the funding and those that were so good that they were ex-ante almost sure to be accepted. Among these proposals a factor, orthogonal to the grade, determined selection. We select the pairs of industry-CZ by considering those with an eligible LabEx satisfying the restriction and estimate the same model as previously. Results are presented in the third section of Table 5. The static coefficient remains of the same magnitude and sign as in the baseline model.

If the marginal factor determining selection, within this group of comparable projects, was not the potential for spillovers, the results above dissipate the concern on selection. To provide further evidence we perform a different exercise and keep only the proposals with a very good grade.[33] In spite of their scientific quality, still 4 of those were rejected, probably due to reasons that were independent of the quality of the proposal itself (for example to ensure some level of geographical distribution across the whole country). For those, given their outstanding scientific quality, it is very unlikely that the anticipation of their local impact on the private sector was the marginal factor used by the jury to determine the selection. Performing the same exercise as above, we show in Table 5 that the results are also unaffected.

5 Channels for spillovers

Section 4 provides robust evidence of sizeable spillovers from the public to the private sector. We both propose a new measure of proximity and provide causal evidence on spillovers. In the current section, we turn to our third main contribution which is to study the mechanisms that fuel these spillovers.

32. Grades between 26 and 32 for proposals filed in 2010 and between 30 and 32 (out of 35) for proposals submitted in 2011, there are 115 such proposals.

33. Greater than 32 out of 35, there are 73 such proposals.

There are three main channels through which spillovers can operate. First, they can result from formal contracting between the public researchers and private firms (*contracting channel*). The second potential channel is mobility of researchers (PhD students and more senior researchers) taking up part-time or full-time positions in the private sector, bringing new ideas to the private firms in the process (*mobility channel*). We include in this category the creation of new startups by public sector researchers. Finally, spillovers can occur through informal contacts between private and public researchers, for instance in the context of outreach events organized by the academic group (*informal channel*).

5.1 Evidence from initial reports

We first exploit a unique source of data, the initial reports written by the funded LabEx shortly after the start of the program, to formalize for the funding agency the governance and potential impacts of the LabEx. These reports have a specific section called "socio-economic impacts of the project", that describe in particular the projected interactions with the private sector, and therefore shed light on the importance of the different channels mentioned above.

As shown in Table 6, that summarizes the content of the reports, 74% of them mention an activity that we characterize as belonging to the contracting channel. Four main types repeatedly appear: contracts (including subcontracting of research by firms), public-private research partnerships, PhD co-supervision and finally licensing agreements of academic patents. Some reports also mention more original types of contracts. For instance, a LabEx specializing in nanotechnologies describes an agreement whereby expensive equipment are provided by private firms in exchange for the sharing of scientific results. The report states that "this type of collaboration can be very fruitful since the lab can obtain state of the art equipment that cannot be otherwise obtained, while the providers of the equipment obtain scientific information that their internal R&D teams cannot obtain."

Table 6 also documents that 30% of reports mention some type of informal contacts. In a number of cases this corresponds to industrial outreach, i.e. the organization of seminars targeted towards industrials and engineers. Some reports in fact use the terminology informal contracts. One LabEx planned a "program of regular meetings between PhD students and researchers with actors of the private sector to build relationships." Finally as mentionned in the introduction, the LabEx ACTION, planned the creation of a "club of partners", described as a "structure of exchange of information between the members of the LabEx and potential partners". They insist on the fact that the membership in this club will not be contingent on a contractual relation with the Labex.

Table 6: Evidence on channels in initial report

Channel	Sub Category	Nb. reports	Share Reports
Contracting		128	74%
	Contracts	78	45%
	Partnerships	46	27%
	PhD co-supervision	15	9%
	Patent licensing	67	39%
Mobility		89	52%
	Startup creation	72	41%
Informal contacts		53	30 %
	Industrial outreach	17	10 %

Notes: This table summarizes the information contained in the initial reports. Column 1 gives the three broad categories of channels (contracting, mobility and informal contacts), column 2 organizes these channels into sub-categories. Column 3 counts the number of reports in each category while column 4 lists the proportion of reports where the category under consideration appears.

Finally, 52% of the reports mention efforts to facilitate the mobility of students and staff to the private sector. Part of this channel corresponds to setting up instruments to facilitate startup creations. It also corresponds to efforts oriented towards helping and encouraging master and PhD students to find a job in the industry. The reports mention "training and exchanging students with the industrial partners" or highlight planned efforts to "create PhD and Postdoc positions at the intersection of different disciplines to create new skills for these young researchers facilitating their professional insertion in the high technology sectors. This will contribute to increase the competitivity of these firms."

5.2 Causal evidence on channels

The picture drawn by the reports is one where all the three channels appear to play a role, the contracting channel being particularly important. These reports provide a comprehensive picture on mechanisms, since they are filed by all funded proposals. They do not allow us however to make causal claims. To show that indeed such mechanisms are at play, we exploit the wealth of data we assembled and provide causal evidence on specific instances of these channels.

5.2.1 The contracting channel

To measure formal contracting, we exploit the data on PhD co-supervision contracts (Cifre program) as well as data on outsourcing from private firms to public research labs recorded in the French research tax credit data. These are two important instances of the contracting channels, though they of course do not capture all the different subcategories mentioned in the reports.

The effect on PhD co-supervision contracts is presented in Figure 2(a). The figure shows the absence of differential pre-trends and a significant increase in these contracts with industries scientifically close to the funded group. The results by quantile presented in Figure B5 in the Appendix, suggest that the effect is concentrated in the top quartile. There is an increase of 5% of the probability of having at least one PhD co-supervision contract in the industry*location pair in the top quartile compared to the bottom one (on average only 6% of industry*location pairs have these type of contracts).

Figure 2(b) shows the effect of the program on R&D outsourcing from private firms. The more exposed industries become more likely to sign at least one outsourcing contract with a public lab. The effect becomes larger a few years after the treatment. The effect also appears concentrated in the top quartile. The results of the static specification (3) are presented in Table 7. On average, doubling the exposure of an industry–location pair, increases PhD co-supervision contracts by 0.39% (base rate of 6%) and outsourcing by 0.2% (base rate of 12%). Outsourcing also increases at the intensive margin, with a doubling in the exposure of an industry–CZ pair implying a 5.6% increase in the amounts outsourced to public labs.[34]

5.2.2 The mobility channel

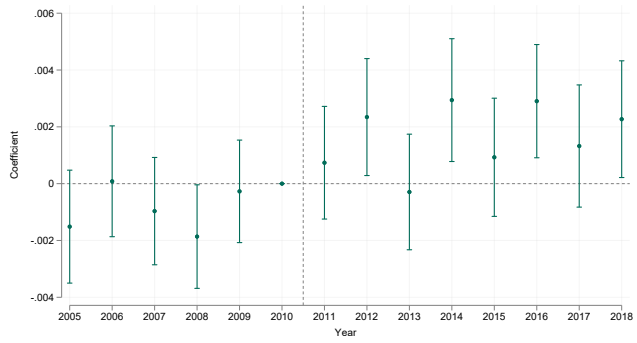
We now turn to the second channel. We can observe mobility by using the administrative data which since 2010 contains movements from the public sector. We have access to the complete French administrative data on employment, so this provides us with a comprehensive measure of movements from the public to the private sector. For any worker in year t we can observe the main occupation of that worker in $t-1$. In particular, we can track public researchers in $t-1$ who have as highest paying job a position in the private sector in t. This is our measure of

34. As a last evidence on this channel, we use the CIS which contains information on collaborations between surveyed firms and universities. While we don't have information on the nature of the university (and in particular can't link them to a specific LabEx or location), we can still look at how the exposure aggregated at the industry level is correlated with changes in the number of collaborations after 2010. The results are presented in Figure B6 in the Appendix, and shows that industries that were exposed more intensely to the LabEx policy witness an increase in the probability to declare a collaboration with a local university.

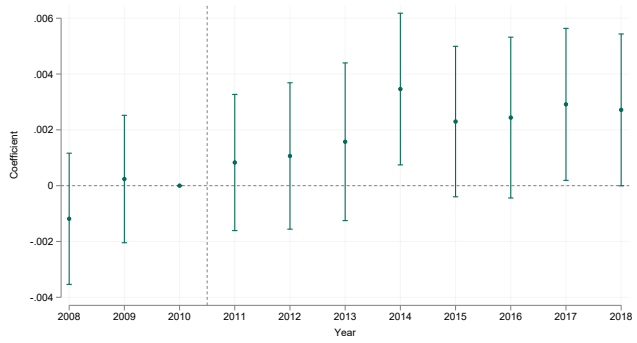
Figure 2: Impact of Labex funding on channels

Contracting channel

(a) PhD Co-supervision

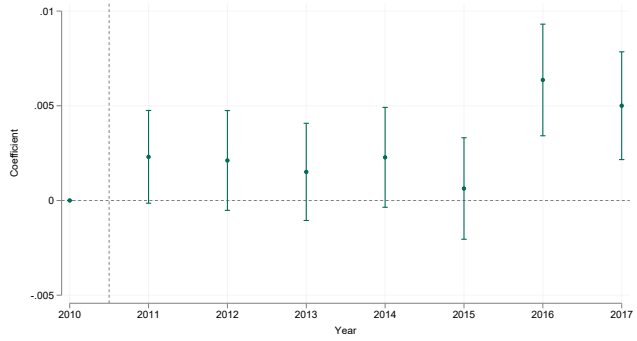


(b) Outsourcing R&D to public labs

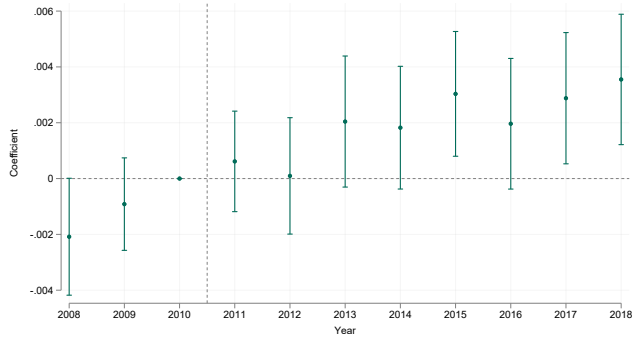


Mobility channel

(c) Researcher transfer



(d) Hiring of young PhD



Notes: These figures are similar to Figure 1(a) but consider the probability that (a) a PhD co-supervision is agreed upon, (b) there is some outsourcing from firms to public labs, (c) there is a transfer of a researcher from the public to the private and (d) some hiring of young PhDs from the public research sector.

mobility from the public research sector to the private sector. Furthermore, we can distinguish junior researchers (PhD students, researchers with temporary teaching contracts) and senior researchers (those that hold permanent research positions) in the administrative data. We can also measure hiring of young PhDs as declared in the tax credit declaration.

The results are presented in Figures 2(c) for total researcher transfers and 2(d) for the hiring of young PhDs. These figures document a significant increase in both categories. Sectors that are closer scientifically to the funded LabEx are more likely to attract public researchers and Phd students from the public sector after the funding shock. Given the constraint on the data that starts recording public sector workers only in 2009, we cannot establish the absence of pre-trends for the total transfer of researchers.

The results of the static specification (3) are presented in Table 7, Panel B. On average, doubling the exposure of an industry-location pair, increases transfers of researchers by 0.30%. The

Table 7: Difference-in-differences estimates of spillover channels

Panel A: Contracting channel			
	Static Coefficient	Obs.	Pre Trends
PhD co-supervision (binary)	0.0033***	59,990 obs (4285 pairs)	-0.0002
	(0.0008)		(0.0009)
Academic Spin-offs (binary)	0.0015***	42,850 obs (4285 pairs)	-0.0000
	(0.0003)		(0.0000)
Outsourcing R&D to public labs (binary)	0.0025***	47,135 obs (4285 pairs)	-0.0005
	(0.0008)		(0.0010)
Outsourcing R&D to public labs (log)	0.0288*	5,183 obs (1031 pairs)	-0.0237
	(0.0155)		(0.0226)
Panel B: Mobility channel			
	Static Coefficient	Obs.	Pre Trends
Transfer of senior academics (binary)	0.0030***	34,280 obs (4285 pairs)	-
	(0.0008)		
Transfer of junior academics (binary)	0.0018*	34,280 obs (4285 pairs)	-
	(0.0012)		
Transfer of researchers (binary)	0.0029***	34,280 obs (4285 pairs)	-
	(0.0011)		
Hiring of young PhDs (binary)	0.0030***	47,135 obs (4285 pairs)	-0.0015*
	(0.0007)		(0.0009)

Notes: Same as Table 2 but using different dependent variables. The absence of pre-trend coefficients for some outcomes are due to the fact that we do not measure them before 2010 (see Section 2).

last line of the Table exploits the information available in the CIR declarations where firms report whether they hired young PhDs. The effect is of the same order of magnitude.

Overall, the initial reports, discussed in Section 5.1, show that all three channels play a role and Section 5.2 validates the finding by providing causal evidence on certain subcategories of these channels. Furthermore, the reports highlight the particular importance of the contracting channel.

6 Policy implications and conclusion

We have shown causal evidence on the existence of spillovers from the public research sector to the private sector. Furthermore we have shown the particular importance of the contracting channel to explain these results. We conclude this paper with a discussion of policy implications of these findings.

The LabEx program we exploit to derive these results is a specific type of public research financing: it targets very high quality research groups deciding to work on a common theme. This type of policy instrument is being more and more widely adopted by European authorities,[35] and universities also increasingly divert funds from traditional discipline based funding to invest in specific themes.[36] This kind of instrument can be particularly well suited to generate spillovers as opposed to individual research grants since they give visibility to the theme locally and also encourage researchers to make particular efforts to share their knowledge. Comparing the different modes of financing public research would be an important topic of future research.

Our results also have policy implications for the financing of private sector innovation. Because of spillovers, whose importance we showed in Section 4.1, the financing of public research can be considered as an indirect policy tool to finance private sector R&D. Another widespread instrument to encourage private sector innovation are tax credit programs, which have been shown to be effective in spurring R&D (Bloom, Van Reenen, and Williams, 2019). As opposed to the financing of public research, such tools can be described as direct instruments of financing, since they directly target R&D spending by firms.

The tax credit program in France is called Crédit d'Impôt Recherche (CIR), and accounts for

35. The *"Exzellenzinitiative"* in Germany, the *"Severo Ochoa"* Centers of Excellence in Spain, the Centers of Excellence in the Nordic countries (descriptive evidence in Möller, Schmidt, and Hornbostel, 2016) or the Initiative d'Excellence in France.

36. There are numerous instances of academic clusters (or centers) of excellence created recently within (or sometimes across) universities such as the University of British Columbia, Stanford University, MIT, or the University of Cambridge.

more than two thirds of direct incentives to innovation for firms (see details in Appendix C.3). Tax credits are earned as a share of reported R&D expenditures, which can be labor or capital costs.[37] A potential limitation of tax credit programs is that they are based on inputs in R&D. Therefore, they cannot target firms that make the most productive use of these inputs. Financing public research on the contrary can benefit only the private firms which are productive in R&D and can thus benefit from spillovers. In that sense this type of financing targets more innovative firms.

While an empirical comparison of the relative impact of these two instruments on private sector innovation is beyond the scope of this paper, we can shed light on the question of what sectors are more affected by these two instruments. To perform this comparison, we allocate the total funding of the two instruments by industry: the LabEx funding is allocated using the sum of exposures over all labs, while the research tax credit is using the claims made by firms.[38]

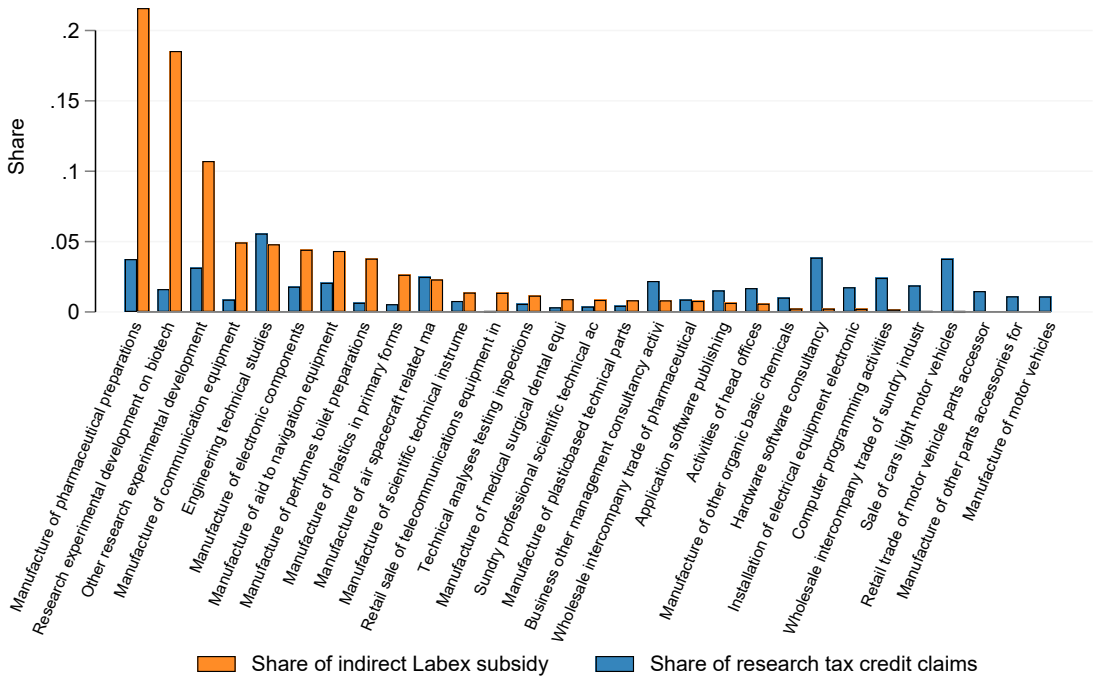
The results are presented in Figure 3. We see that the distribution of exposure to the LabEx program is much more skewed towards a few sectors. The two sectors that benefit the most are the scientific R&D (almost 30%) and the pharmaceutical sector (more than 20%).[39] On the contrary the benefits of the CIR are much more evenly distributed across sectors, including in sectors where innovation should not be central, such as computer consultancy. This evidence is coherent with the idea that financing public research to spur private sector innovation might be a better instrument to target truly innovative firms than more direct tools such as tax credit programs. Comparing these instruments and understanding their complementarities is an important avenue for future research.

37. One of the peculiarities of the French research tax credit is that it has a very high ceiling (which implies a drop from a 30% to a 5% rate), whereas many similar programs in other countries only apply to SMEs.

38. As noted in the data section, we define a firm's industry as the industry of the largest non financial unit, to avoid attributing a large weight to finance because of holding companies.

39. This justifies their exclusion in robustness checks of our main analysis (Table 4)

Figure 3: Share of benefits by sector from LabEx vs CIR programs



References

Abramovsky, Laura, Rupert Harrison, and Helen Simpson. 2007. "University research and the location of business R&D." *The Economic Journal* 117 (519): C114–C141.

Abramovsky, Laura, and Helen Simpson. 2011. "Geographic proximity and firm–university innovation linkages: evidence from Great Britain." *Journal of economic geography* 11 (6): 949–977.

Agrawal, Ajay, and Rebecca Henderson. 2002. "Putting patents in context: Exploring knowledge transfer from MIT." *Management science* 48 (1): 44–60.

Akcigit, Ufuk, Douglas Hanley, and Nicolas Serrano-Velarde. 2021. "Back to Basics: Basic Research Spillovers, Innovation Policy, and Growth." *Review of Economic Studies* 88 (1): 1–43.

Ankrah, Samuel, and AL-Tabbaa Omar. 2015. "Universities–industry collaboration: A systematic review." *Scandinavian Journal of Management* 31 (3): 387–408.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2021. "Knowledge Spillovers and Corporate Investment in Scientific Research." *American Economic Review* 111 (3): 871–98.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. 2019. "Public R&D investments and private-sector patenting: evidence from NIH funding rules." *The Review of economic studies* 86 (1): 117–152.

Bergeaud, Antonin, Arthur Guillouzouic, Emeric Henry, and Clément Malgouyres. 2022. "Proximity of Firms to Scientific Production." *mimeo.*

Bikard, Michaël, and Matt Marx. 2020. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science* 66 (8): 3425–3443.

Bloom, Nicholas, John Van Reenen, and Heidi Williams. 2019. "A Toolkit of Policies to Promote Innovation." *Journal of Economic Perspectives* 33 (3): 163–84.

Borusyak, Kirill, and Peter Hull. 2021. *Non-random exposure to exogenous shocks.* Technical report w27845. National Bureau of Economic Research.

Carayol, Nicolas, Emeric Henry, and Marianne Lanoë. 2020. "Stimulating Peer Effects? Evidence from a Research Cluster Policy."

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science* 48 (1): 1–23.

Cristelli, Gabriele, Gaetan de Rassenfosse, Kyle Higham, and Cyril Verluise. 2020. "The Missing 15 Percent of Patent Citations." Available at SSRN 3754772.

Cross, Samuel, Yeanuk Rho, Henna Reddy, Toby Pepperrell, Florence Rodgers, Rhiannon Osborne, Ayolola Eni-Olotu, Rishi Banerjee, Sabrina Wimmer, and Sarai Keestra. 2021. "Who funded the research behind the Oxford–AstraZeneca COVID-19 vaccine?" *BMJ Global Health* 6.

Dahl, Michael S, and Christian ØR Pedersen. 2004. "Knowledge flows through informal contacts in industrial clusters: myth or reality?" *Research policy* 33 (10): 1673–1686.

De Fuentes, Claudia, and Gabriela Dutrénit. 2012. "Best channels of academia–industry interaction for long-term benefit." *Research Policy* 41 (9): 1666–1682.

Fernandes, Cristina I, and João JM Ferreira. 2013. "Knowledge spillovers: cooperation between universities and KIBS." *R&D Management* 43 (5): 461–472.

Guillouzouic, Arthur, and Clément Malgouyres. 2020. *Évaluation des effets du dispositif CIFRE sur les entreprises et les doctorants participants.* Technical report.

Hausman, Naomi. 2021. "University Innovation and Local Economic Growth." *The Review of Economics and Statistics* (March): 1–46.

Howell, Sabrina T. 2017. "Financing innovation: Evidence from R&D grants." *American Economic Review* 107 (4): 1136–64.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science." *Quarterly Journal of Economics* 133 (2): 927–991.

Jaffe, Adam B. 1989. "Real effects of academic research." *The American economic review,* 957–970.

Kantor, Shawn, and Alexander Whalley. 2014. "Knowledge spillovers from research universities: evidence from endowment value shocks." *Review of Economics and Statistics* 96 (1): 171–188.

———. 2022. "Moonshot: Public R&D and Growth." *mimeo.*

Kiszewski, Anthony, Ekaterina Galkina Cleary, Matthew Jackson, and Fred Ledley. 2021. "NIH funding for vaccine readiness before the COVID-19 pandemic." *Vaccine* 39:2458–2466.

Maraut, Stéphane, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec. 2008. *The OECD REGPAT Database: A Presentation.* Working Paper 2008/02. OECD Science, Technology and Industry.

Möller, Torger, Marion Schmidt, and Stefan Hornbostel. 2016. "Assessing the effects of the German Excellence Initiative with bibliometric methods." *Scientometrics* 109 (3): 2217–2239.

Moretti, Enrico, Claudia Steinwender, and John Van Reenen. 2019. *The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers.* Technical report. National Bureau of Economic Research.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German-Jewish Émigrés and U.S. Invention." *American Economic Review* 104 (10): 3222–3255.

Myers, Kyle R, and Lauren Lanahan. 2022. "Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy." *American Economic Review* 112 (7): 2393–2423.

Perkmann, Markus, Valentina Tartari, Maureen McKelvey, Erkko Autio, Anders Broström, Pablo D'este, Riccardo Fini, Aldo Geuna, Rosa Grimaldi, Alan Hughes, et al. 2013. "Academic engagement and commercialisation: A review of the literature on university–industry relations." *Research policy* 42 (2): 423–442.

Srebrovic, Rob. 2019. "Expanding your patent set with ML and BigQuery." Google Cloud Data Analytics https://cloud.google.com/blog/products/data-analytics/expanding-your-patent-set-with-ml-and-bigquery.

Teichgraeber, Andreas, and John Van Reenen. 2022. *A policy toolkit to increase research and innovation in the European Union.* Discussion Paper DP1832. Centre for Economic Performance.

Waldinger, Fabian. 2012. "Peer Effects in Science – Evidence from the Dismissal of Scientists in Nazi Germany." *The Review of Economic Studies* 79 (2): 838–861.

Online Appendix

A Additional tables

Table A1: Summary statistics on the baseline estimation sample

Panel A. Sector × CZ-level statistics			
R&D and exposure variables	Mean	p50	p90
Exposure (in milions euros)	0.19	0.00	0.15
Proximity	0.05	0.00	0.07
Total employment	575.55	135.54	1029.41
Engineer employment	108.96	6.12	111.24
# plants	50.88	8.00	97.00
# plants employing engineers	3.39	0.00	4.00
# patents	1.09	0.00	0.00
Cooperation variables			
Outsourcing R&D to public sector	0.12	0.00	1.00
PhD co-supervision	0.05	0.00	0.00
Transfer of researchers	0.09	0.00	0.00
Hiring of young PhDs	0.07	0.00	0.00
Academic spin-off	0.01	0.00	0.00
Observations: 59,990 — # (CZ × NAF) : 4,285			
Panel B. CZ-level statistics			
# of 5-digit sector	8.05	8.00	12.00
Total employment (in 1000s)	64.62	30.13	89.23
# plants	5737.43	2848.50	8906.00
# plants employing engineers	381.83	152.00	612.00
Observations: 532 — # CZ : 38			

Notes: This Table presents descriptive statistics regarding the estimating sample.

Table A2: Probability to be funded according to LabEx grades and characteristics

Dep. var. : P(funded)	
Grade: Team quality	0.0379
	(0.050)
Grade: Project's scientific ambition	0.1231***
	(0.036)
Grade: Innovation and impact	0.0610
	(0.038)
Grade: Teaching quality	-0.0016
	(0.039)
Grade: Management quality	-0.0249
	(0.038)
Grade: Partner univ. joint strategy	0.0329
	(0.041)
Grade: Adequation ambition / funding	0.1173***
	(0.035)
Second wave	-0.2758***
	(0.046)
Funding requested (log)	0.0707***
	(0.026)
R^2	0.384
Observations	340

Notes: This Table presents the results of an OLS regression of a dummy indicating if a lab received funding on the grades it obtained over the seven dimensions of grading and basic characteristics (year of application and amount of funding requested).

Table A3: Origin of engineers

	Static Coefficient	Obs.	Pre Trends	Init. share
R&D hours	0.0091**	42,560 obs (4285 pairs)	-0.0022	1.000
	(0.0036)		(0.0025)	
Incumbent R&D hours	0.0090**	42,130 obs (4285 pairs)	-0.0015	0.868
	(0.0036)		(0.0027)	
Industry Movers R&D hours	-0.0004	38,300 obs (4285 pairs)	-0.0069	0.074
	(0.0091)		(0.0128)	
...incl. Ind. & CZ Movers R&D hours	-0.0051	37,382 obs (4285 pairs)	-0.0098	0.065
	(0.0091)		(0.0128)	
CZ Movers R&D hours	0.0186***	35,488 obs (4285 pairs)	0.0063	0.038
	(0.0053)		(0.0044)	
Entrants R&D hours	0.0236***	35,862 obs (4285 pairs)	-0.0287	0.029
	(0.0072)		(0.0063)	

Notes: Each line corresponds to a different dependent variable decomposing total hours worked by engineers in year N. Coefficients are obtained running a pseudo-Poisson maximum-likelihood static difference-in-differences specification over the period 2009-2018 (see model (3)). The reported coefficient corresponds to the exposure variable interacted with a post (i.e. after 2010) dummy variable. The penultimate column shows the average value of the pre-trend coefficients of the model (4), estimated over the same period 2005-2018. The last column shows the share of each category in the total hours of engineers in 2010. The unit of observation is a pair of commuting zone $\times$ 5-digits industry. All models include a commuting zone-industry fixed effects and a set of commuting zone-year dummies (the last model also include a set of 5 digit industry-year fixed effects). Standard errors are clustered at the pair level.

Table A4: Removing generalist journals and aggregating journals by field

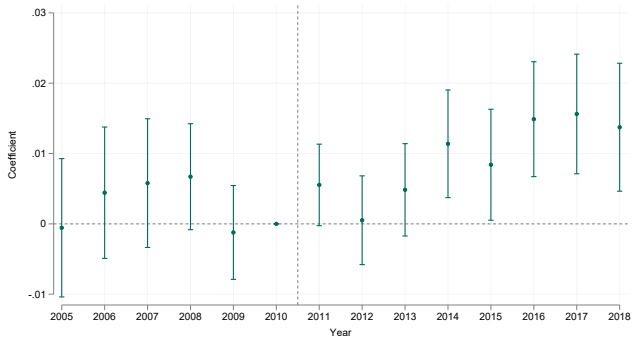
	Static Coefficient	Obs.	Pre Trends
1. Remove generalist journals (Herf $<$ 0.2)	0.0081**	47,986 obs (4285 pairs)	0.0012
	(0.0032)		(0.0031)
2. Remove generalist journals (Herf $<$ 0.5)	0.0087***	47,986 obs (4285 pairs)	0.0004
	(0.0033)		(0.0031)
3. Aggregate journals by field	0.0074*	47,986 obs (4285 pairs)	-0.0045
	(0.0041)		(0.0043)

Notes: This Table replicates Table 2 (row 1.) but changes the measure of exposure by changing the set of journals considered (see Section 3.1). Line 1. remove journals which are too generalist as measured by an Herfindahl index of their Crossref broad scientific fields (18 categories) lower than 0.2. Line 2. does the same but uses a threshold of 0.5. Line 3. aggregates journals in about 200 categories based on the Crossref detailed scientific fields.

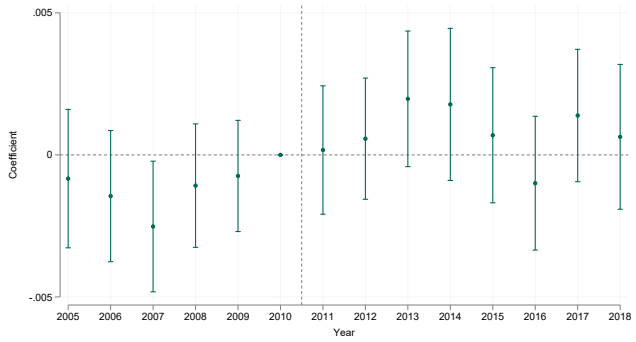
B Additional figures

Figure B1: Impact of Labex funding on R&D effort

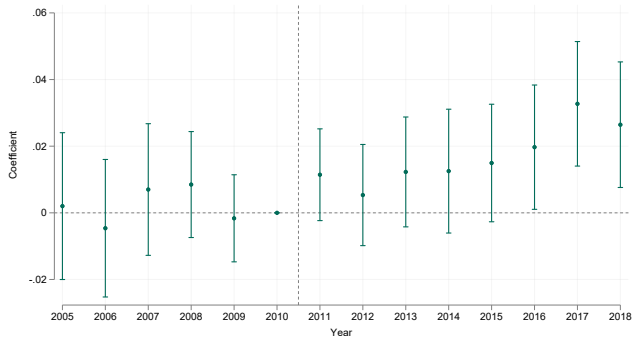
(a) R&D hours (log)



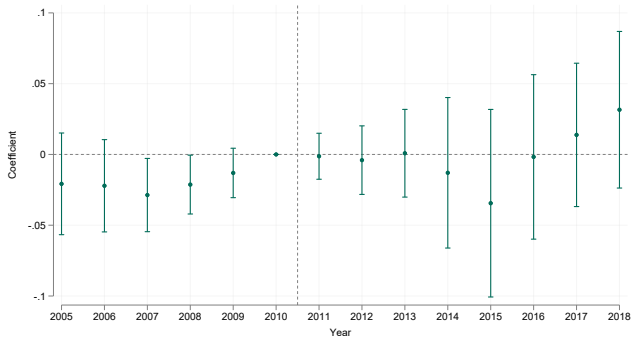
(b) R&D hourly wage (log)



(c) Total R&D claims (log)

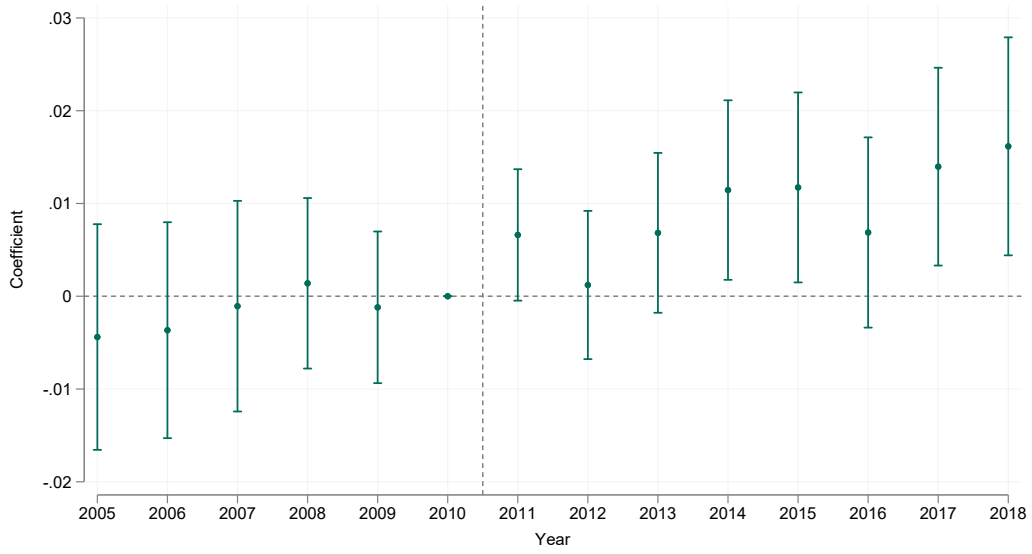


(d) Number of patents



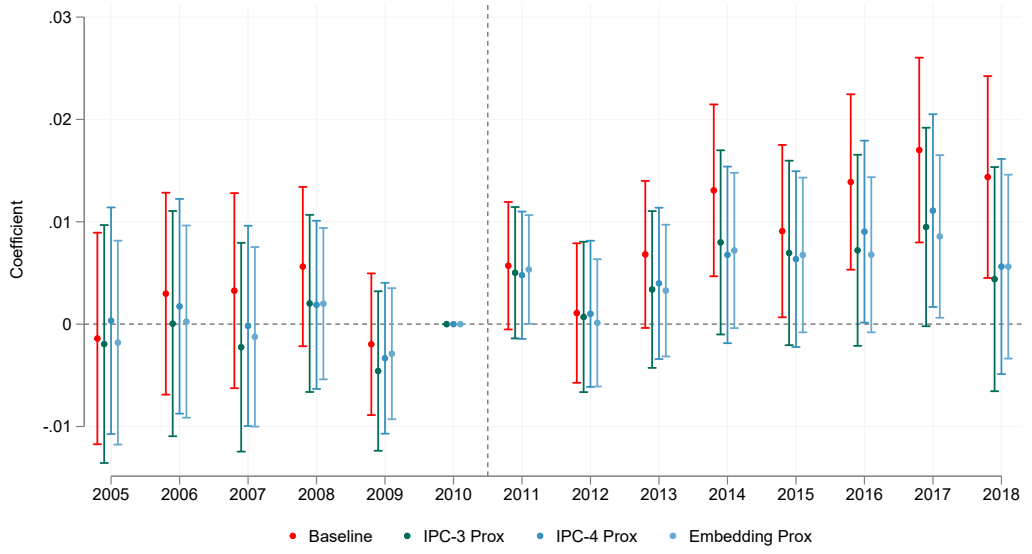
Notes: These figures are similar to Figure 1(a) but consider alternative dependent variables.

Figure B2: Impact of Labex funding on R&D hours controlling for industry–time effects



Notes: see Figure 1. The regression now includes 2 digits industry–year fixed effects.

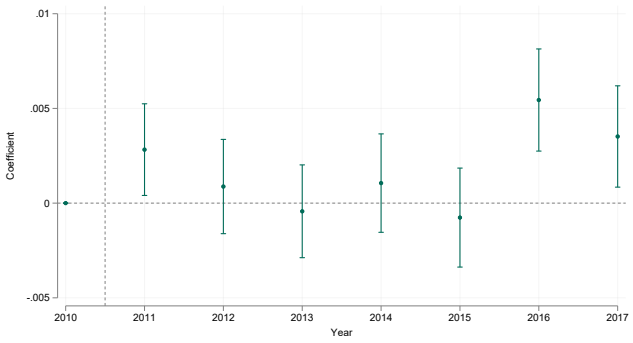
Figure B3: Impact of the Labex funding on R&D wage bill with different proximity measures



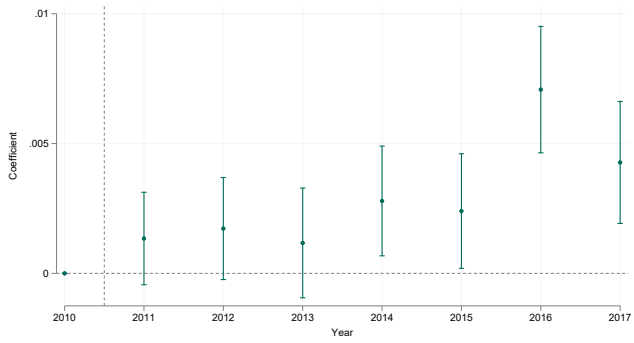
Notes: These figures reproduce Figure 1(a) using alternative measures of proximity to build exposure, as described in Section 3.

Figure B4: Impact of Labex funding on channels

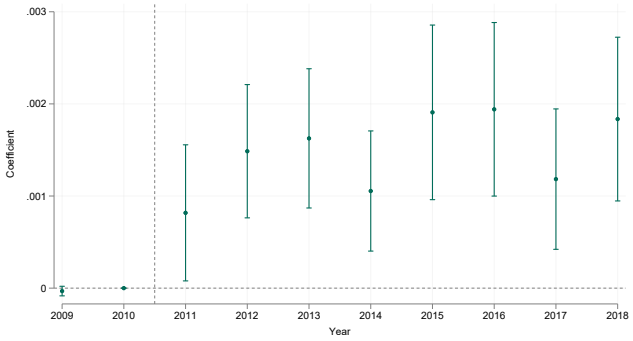
(a) Transfer of junior academics (binary)



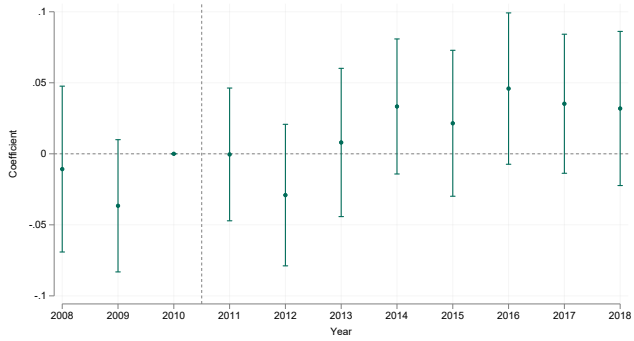
(b) Transfer of senior academics (binary)



(c) Academic spin-offs (binary)



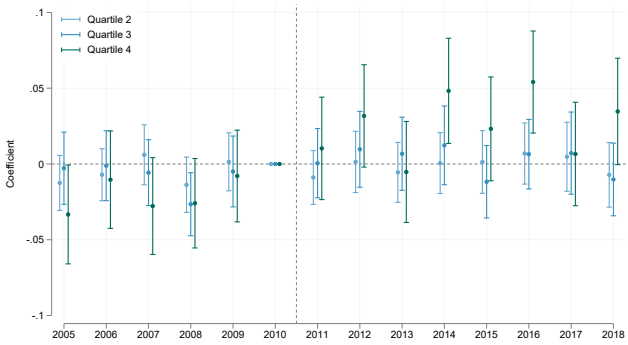
(d) Outsourcing R&D to public labs (log)



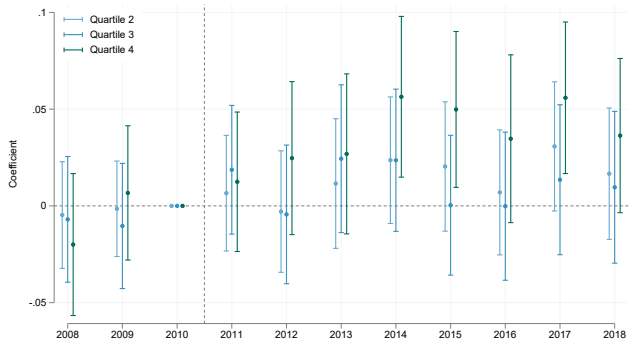
Notes: These figures reproduce Figure 1(a) for alternative dependent variables.

Figure B5: Impact of Labex funding on channels (binary variables) by quantile

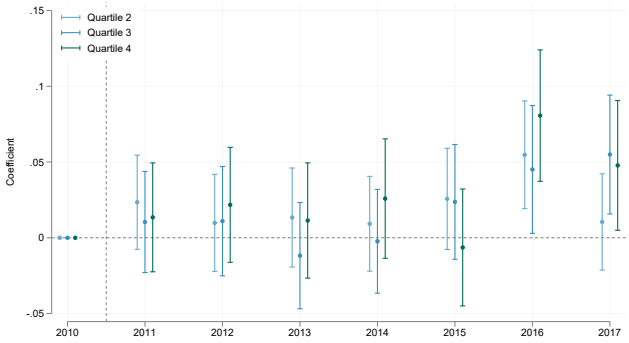
(a) PhD Co-supervision



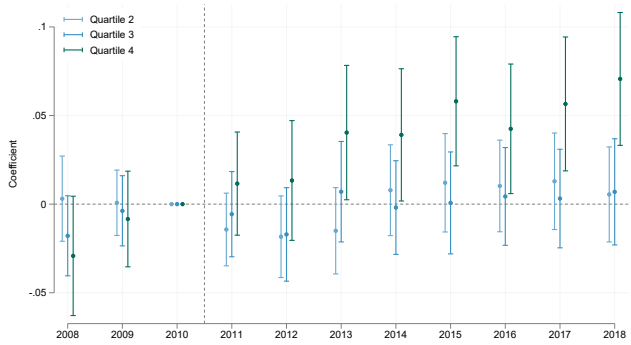
(b) Outsourcing R&D to public labs



(c) Transfer of researchers

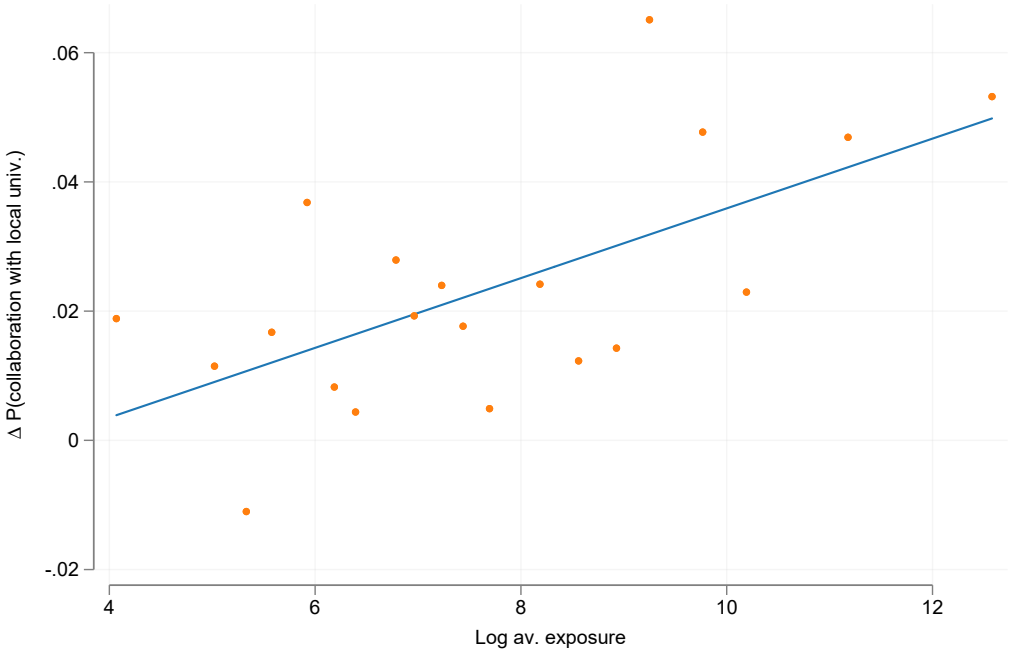


(d) Hiring of young PhDs



Notes: These figures reproduce Figure 1(b) for alternative dependent variables.

Figure B6: Change in the probability to collaborate with a university versus log of av. exposure to LabEx



Notes: This figure shows the variation in the probability to declare collaborating with a local university for firms in an industry surveyed in the waves 2004, 2008 and 2010 relative to firms in that industry surveyed in the waves 2012, 2014 and 2016 in the CIS Survey. This change is plotted against the log of the average exposure to the Labex policy of an industry (across commuting zones).

C Data sources and variable construction

C.1 Patent data

Patent–firm identifier matching procedure The matching procedure of French patents to the firm identifier (Siren number) of their assignees is implemented in several steps.

We rely on the harmonized PSN (Patstat Standardized Name) identification of assignees and inventors in the Patstat Spring 2020 Database. We select all such identifiers whose country code is recorded as being France at least once.[40] We further require that the type of applicant recorded in Patstat is not specified as being "individual" and is not missing, so as to focus on companies and public organizations. We complement this list with a matching between firm identifiers and French patent applicants available in the scanR search engine. This selection leaves us with 76,582 PSN identifiers to match with French legal unit identifiers. Thanks to this already existing matching, 34,150 identifiers are matched with a Siren firm identifier from the start.

For those that do not match, we proceed in a number of additional steps. In a first step, we try to match directly the names recorded in the Patstat database with firm names. We match names in priority with sources including primarily innovative firms (INPI patents including the Siren identifier, firms that obtain research tax credit, firms in the R&D survey), and then match to firms with no name duplicates in the Sirene (all legal units) registry. This step complements our list of French PSN identifiers in Patstat with 2,313 Siren identifiers.

In a second step, we use fuzzy matching techniques (using both the reclink stata package and Jaro-Winkler distances) on names of firms found in the above-mentioned databases of innovative firms, which are a priori very susceptible of applying for patents. We match only on this very limited set of firms because fuzzy matching procedures with the universe of firms in the registry would be both computationally costly and lead to potentially high rates of type I errors. This step adds 4,105 new siren identifiers to our list of applicants.

Next, we send requests of patent applicant names to a major online search engine conditioning on web domains which contain historical registries of French legal units[41]. This allows for a fuzzy matching, where the search engine is able to find the underlying company even though some parts of its name make it difficult to match through fuzzy matching techniques (for instance the presence of very common words which add little value but many characters, making the string distance very high but which the search engine easily ignores). This steps adds 4,910 new siren identifiers.

40. psn_id groups several person_id, and the country code is recorded at the person_id level. This means that, for instance for multinational companies, we keep the psn_id which correspond to at least one person identifier located in France.

41. For instance societe.com.

As a last step, we take the list of yet unmatched names and use the web "batch search" tool of Bureau van Dijk's Orbis database, which has a built-in module of fuzzy matching of company names with BvD identifiers (Orbis identifier), which can be directly converted into a Siren. This final step adds 2,151 siren identifiers.

Finally, we consolidate our identifiers by name and by psn_id, and manually search for the siren identifiers of the largest unmatched applicants.[42] We end up with 47,545 psn_id which are associated with 33,826 different siren identifiers, associated with approximately 963 thousand patent applications.

Details on PatCit data Text mining methods are applied to both the dedicated section of the application, and to the text describing the invention, in order to extract bibliographic references.[43] In the case of citations to the NPL, the vast majority of which are academic articles, the database collects the DOIs (digital object identifiers) of the cited publications, and thus enables a match with other bibliographic databases. The database includes 27 million academic bibliographic references, of which more than 11 million could be associated with a DOI. The database is described in more detail in Cristelli et al. (2020).

C.2 Linked employer-employee data

R&D wage bill In the main text, we use as our measure of spending on R&D labor, the wages reported in the administrative data DADS for engineering occupations. We use positions with an occupation and socio-professional category (PCS) beginning with 38: "Engineers and technical managers of companies".[44]

This measure has the advantage of covering the entire private sector over a long period of time. In this appendix, we validate our measure by comparing it to R&D employment as measured by the R&D survey (RDS), for the subsample of firms present in both databases. We compare the wage bill of engineers in the DADS to the wage bill of R&D personnel in the RDS. In order to account for the size of the firm, we normalize the relevant wage bill by total sales.

In Table C1, we compare these variables of interest between the two datasets for all firms (legal units) present in each of the two sources continuously over the period 2009–2016 in the first line. In the second line, we restrict to the sample of firms reporting a positive value in both

42. These are often firms which have changed their name over time, which we associate with their current identifier to obtain a consistent patenting history if we were to use older periods.

43. The database is available at https://cverluise.github.io/PatCit/

44. The advantage of this definition is that there is no break in the series and that it is therefore available over the entire study period, before and after the 2008 reform. Before 2009, it is not mandatory to provide this detailed PCS for companies with less than 20 employees. Nevertheless, there was a break in the series in 2009, even for companies with more than 20 employees, making it impossible to consider the long series for this detailed variable reliable.

sources. The correlation is 72.4 % in the whole sample and rises to 84.5 % in the sub-sample of firms where the variable takes a positive value. In the next column we compute the difference between the two measures. The difference is close to 0 at the median but negative on average, meaning that both the engineer measure and the employee measure of R&D in the DADS tend to underestimate actual R&D personnel spending. This can be explained by the fact that the definition of employees contributing to R&D (for example, as defined by the CIR, the French tax credit) is broader than the definition of employees' positions as R&D-oriented, and that the fact that not all the company's engineers are R&D-oriented is not sufficient to make up for this difference.

Table C1: Comparison of the wage bill of R&D staff (in the R&D survey) and the engineer (in the DADS)

Variable	Sample	N	ρ	Gap: mean	p50	p10	p90
DADS Engineer	Full	52,525	0.724	-0.178	0.002	-0.300	0.125
DADS Engineer	Positive var.	45,106	0.845	-0.112	0.007	-0.202	0.147

NOTES : ρ = correlation coefficient. Gap $:= \frac{\text{Engineer wage bill DADS}}{\text{Sales}} - \frac{\text{R\&D wage bill RDS}}{\text{Sales}}$, where "R&D wage bill RDS" refers to wage R&D expenditure in the R&D survey (RDS). The "positive var." sample concerns companies reporting a positive amount of payroll in the DADS as engineers.

R&D plants We use a similar procedure, exploiting the administrative data DADS, to identify R&D intensive plants. This is not achievable using surveys on R&D which are administered at the firm level. We define an R&D plant as an establishment with more than 20% of its wage bill spent on R&D wages (as defined above). We use this information to compute the variable *Hours in R&D plants*, defined as the total number of hours worked in the R&D plants (defined above).

Labor mobility Since 2009 the DADS include public sector employees. This allows us to measure mobility from the research public sector to the private sector from 2010 onward. We define a mover as a worker whose main job in $t-1$ was in the public sector in a research occupation and in t, gets most of her salary from the private sector. We distinguish junior movers (those who were PhDs or in teaching postdoc positions in $t-1$) and senior movers (those who had a permanent position in research in the public sector in $t-1$).

C.3 Research tax credit

Description of the program The French research tax credit program (Crédit Impot Recherche CIR) was set up in 1983. Any firm, including large ones, can participate. The eligible spending covers R&D related expenditures, including wages, investments and subcontracting. The credit is equal to 30% of the spending when the spending is less than 100 million euros, and 5% above. In 2019 more than 7 billion euros were spent on CIR with 26 900 firms making claims.

Data While the MVC CIR database contains only the total amount of tax credit but measured over a long time period (2000-2018), the GECIR database is available only from 2008 onward (so that we cannot observe pre-trends), but features a detailed breakdown of R&D expenditures eligible for the tax credit. In particular, we exploit a variable indicating the amount of R&D outsourced to public organizations (*Outsourcing R&D to public labs*), as well as the amounts of wages paid to young PhD graduates (*Hiring of young PhDs*). These pieces of information are very well recorded since they are used in the calculation of the tax credit. Because these are relatively rare outcomes, we allocate them to the CZ where the firm has the largest share of engineers, rather than splitting them according to this share, as for the overall RTC claims.

C.4 PhD co-supervision

Description of the program The Cifre program is a program, set up in the early 2000s to encourage contacts between public research labs and the industry. The candidate firm and public lab have to submit an application to the national agency (ANRT) and if selected receive a subsidy. The student typically shares her time between the two partners. In 2018 there were around 1500 Cifre contracts signed per year.[45]

Data We obtained data on all Cifre contracts at the individual level, where we can identify the collaborating company with the national firm identifier, the municipality where the PhD student is employed, the public research lab co-supervising the student, the statutory wage, and the date when the 3-years contract starts. These data are available from 2003 to 2018. The variable *PhD co-supervision* we construct in this way is used in the analysis of mechanisms.

C.5 Academic spin-offs

Description of the program The JEU (*Jeunes entreprises universitaires*) program targets academic spinoffs. Qualifying firms need to be launched by students or faculty members in universities, who need to hold at least 10% of the capital. Beneficiaries are young (less than 11 years old), SME (less than 250 employees), with a high R&D intensity. Firms that qualify get reductions in corporate tax rate as well as payroll exemptions for workers related to R&D.

Data We obtained data on firms registered as JEU. These data allow us to build the variable *academic spinoffs*, used as outcome and to illustrate channels. The JEU program was launched in 2009, and only few firms benefited from it in the first two years, so that we mostly observe the outcome concomitantly with our funding shock.

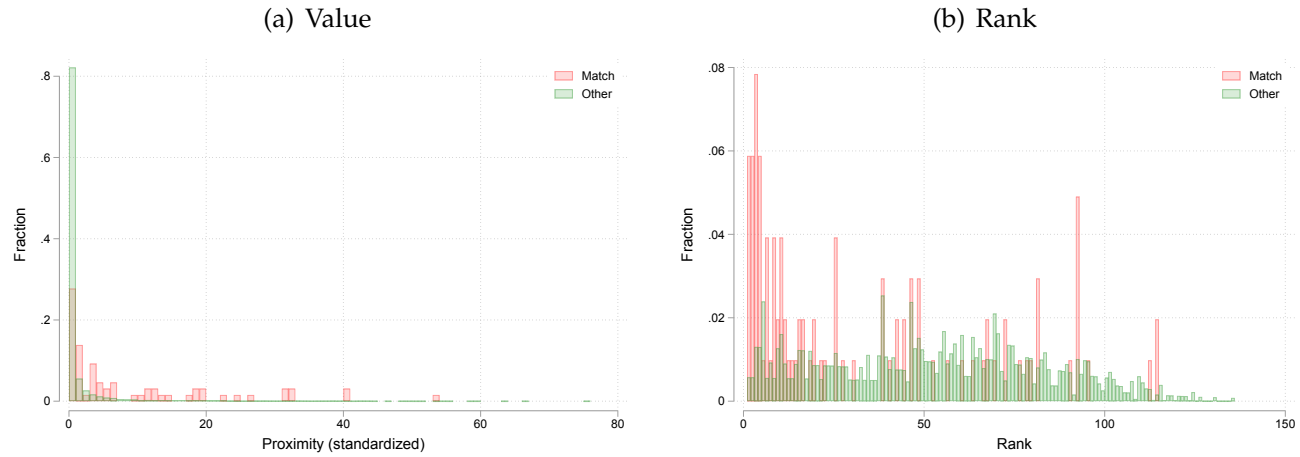
45. See Guillouzouic and Malgouyres (2020) for a complete description of the program.

D Validation of the proximity measure

We provide some validation of our novel indicator of proximity introduced in Section 3.1.

We first exploit the initial reports (see Section 5.1 for details). These reports mention potential collaborations with firms. For instance the LabEx ACTION, mentioned in the introduction, mentions a number of firms by name that it describes as those that could be "interested by the research activities of the LabEx". The report identifies them as potential members of the club of partners. We thus hand-collected all the instances where firms were mentioned in these reports and matched each firm with its sector. From this, we compute a number of matches between a given sector and a given LabEx. We show in Figure D1 that our measure of proximity is a predictor of whether an industry is mentioned. In Panel (a), we show that matched firms are much less likely to appear as having zero proximity than unmatched ones, and that the distribution is shifted towards higher values. In Panel (b) we show that matched firms appear very often among the five closest industries, while unmatched firms have a fairly uniform distribution of ranks.

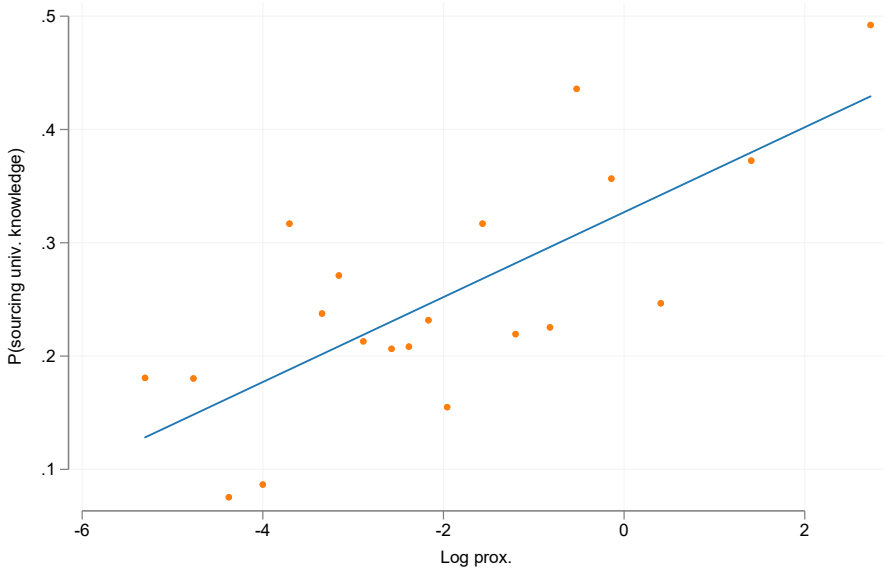
Figure D1: Proximity for matched and unmatched firms



Notes: This figure compares firms that match the firms directly quoted in the LabEx initial reports, and firms that do not. Panel a shows the distribution of a standardized value of proximity for both groups. Panel b shows the distribution of proximity ranks for both groups.

Our second exercise uses three vintages of the Community Innovation Survey (CIS): 2004, 2006 and 2010. In each of these waves, firms are asked to what extent they source their knowledge from universities (0: not at all to 3: a lot). We calculate for each sector the share of firms that don't answer 0 to this question and look at the correlation between this share and the sum of proximities taken across all LabEx (in log). Figure D2 shows a binned scatterplot of the probability for firms of an industry to source knowledge from universities, plotted against the log of the sum of LabEx proximities. It shows a clear positive correlation between both variables, which further supports the fact that our proximity variable captures well the existing proximity between firms and universities.

Figure D2: Probability to use university knowledge versus industry–LabEx proximity



Notes: This figure presents a binned scatterplot of the average probability to use knowledge produced by a university by firms surveyed in waves 2004, 2006 and 2010 of the CIS in an industry, plotted against the (log of the) sum of LabEx proximities.

E Alternative geographical spillovers

Our baseline measure of proximity assumes that there are no spillovers across different CZ. In this section, we explore how our main results are affected when we relax this assumption. We use two approaches. First, we calculate a "neighboring CZ exposure", an exposure that is based on the funding received by LabEx in adjacent CZ and include it in our model. Second, we distribute the funding received by a given LabEx in all CZ with an exponential decay based on distance. This extension is particularly relevant for LabEx which are located near the border of a CZ and are thus likely to impact firms located across this border.

Neighboring CZ. We first define a measure of exposure based on neighboring CZ as follows:

$$\text{expo}_{ik}^{N} = \sum_{l \in N(k)} d_l \cdot \text{prox}_{li},$$

where $N(k)$ denotes the set of CZ that are adjoining CZ k and prox_{li} is the same as in the baseline (1). This expo_{ik}^{N} captures the potential spillover from neighboring CZ and we include it as a control variable in our static and dynamic models (respectively equations (3) and (4)) which become:

$$Y_{ikt} = \mathbb{1}\{t > 2010\} \times \left(\beta \ln\left(1 + \text{expo}_{ik}\right) + \gamma \ln\left(1 + \text{expo}_{ik}^{N}\right)\right) + \alpha_{ik} + \delta_{tk} + \varepsilon_{ikt}$$

and

$$Y_{ikt} = \sum_{\substack{d=2005 \\ d \neq 2010}}^{2017} \left(\mathbb{1}\{t = d\} \times \left(\beta_d \ln\left(1 + \text{expo}_{ik}\right) + \gamma_d \ln\left(1 + \text{expo}_{ik}^{N}\right)\right)\right) + \alpha_{ik} + \delta_{t,k} + \varepsilon_{ikt}$$

Results are presented in Table E1, line 1, using the same sample as in the baseline. They warrant our assumption that spillovers are mostly concentrated with a Commuting Zone and that the control group made of CZ with no (accepted) LabEx is essentially not affected by the treatment.[46]

Continuous distance. Mainland France counts more than 35,000 municipalities which constitute a very fine grid of the territory. We use this to calculate a measure of exposure for each CZ, including those without any LabEx. Formally, let $c \in \mathcal{C}_k$ denotes a given city in CZ k and c_l the

46. To pursue the analysis further, one possibility is to construct a measure of exposure that does not take into account geographical border. That is, the sum is taken over all k in equation (2). Using this as another control in our models yields results that are consistent with Line 1 of Table E1.

city where LabEx l is located. Then we can define weights as:

$$\omega_{k,l} = \bar{v}_l \sum_{c \in \mathcal{C}_k} e^{-\nu \delta(c,c_l)},$$

where $\delta(c, c_l)$ denotes the distance (in km) between two cities c and c_l, ν is a depreciation parameter and $\bar{v}_l$ ensures that the weights sum to one for each LabEx:

$$\bar{v}_l = \frac{1}{\sum_k \sum_{c \in \mathcal{C}_k} e^{-\nu \delta(c,c_l)}}.$$

Then, the continuous measure of exposure is defined as:

$$\text{expo}_{ik}^N = \sum_l \omega_{k,l} d_l \cdot \text{prox}_{li}.$$

We then need to set a value for ν. The distance at which half of the spillover has faded away is equal to $\log(2)/\nu$. We set the value of ν such that this distance is equal to 10km and also show results when this value is set to 5 and 50km respectively. All of this is presented in Table E1, lines 2 to 4. To get a sense on the geographical distribution of spillovers, we plot different quantities in Figure E1 (see also Figure E2 for comparison with the measures used in the core of the paper). First, we report the value of:

$$\sum_l d_l \bar{v}_l e^{-\nu \delta(c,c_l)},$$

at the city level, with ν taken equal to 5, 10 and 50 respectively. Second, we plot the value of the aggregate exposure by CZ:

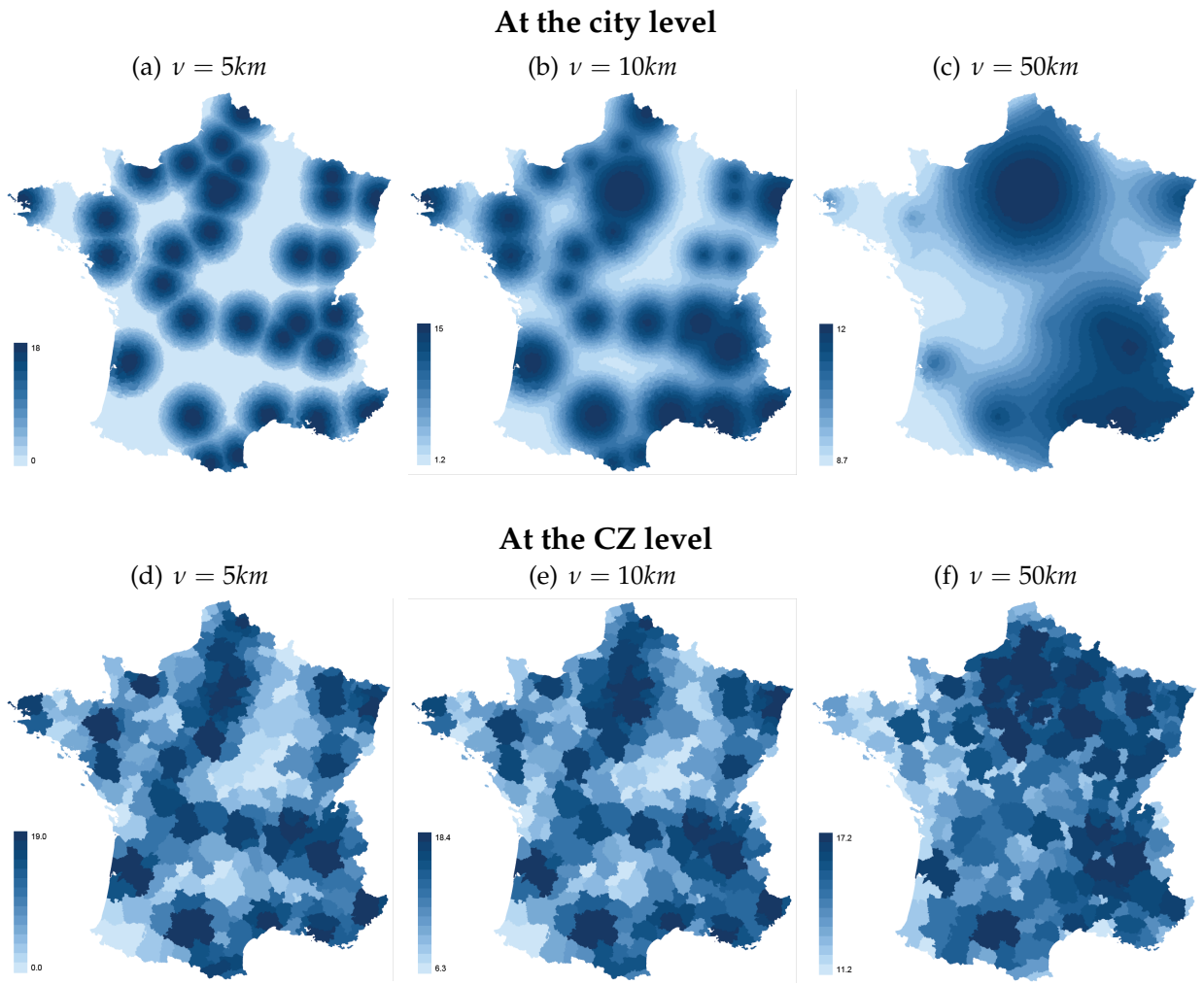
$$\sum_i \sum_l \omega_{k,l} d_l prox_{li}.$$

Table E1: Robustness checks - geographical spillover

	Static Coefficient	Obs.	Pre Trends
Baseline	0.0093**	42,301 obs (3761 pairs)	0.0013
	(0.0037)		(0.0042)
1. Neighboring CZ shock	0.0002	48,138 obs (4308 pairs)	-0.0011
	(0.0046)		(0.0021)
Baseline shock	0.0090***		0.0015
	(0.0035)		(0.0032)
2. Continuous distance shock (10km)	0.0104***	170,871 obs (19,582 pairs)	0.0005
	(0.0023)		(0.0023)
3. Continuous distance shock (5km)	0.0099***	170,871 obs (19,582 pairs)	0.0007
	(0.0024)		(0.0024)
4. Continuous distance shock (50km)	0.0146***	170,871 obs (19,582 pairs)	0.0006
	(0.0029)		(0.0031)

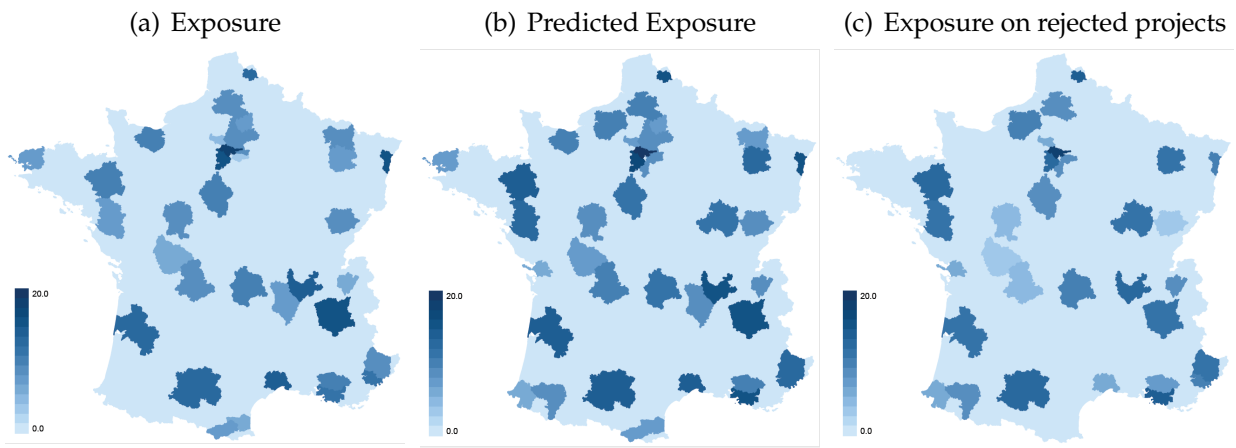
Notes: This Table presents the results of the same estimation as in Table 2, using as dependent variable the log of the total wage bill of engineers and alternative shocks accounting for broader geographical spillovers (see Section E).

Figure E1: Geographical spillovers



Notes: These maps report the value of $\sum_l d_l \bar{v}_l e^{-\nu\delta(c,c_l)}$ for each city (first line) for ν respectively set to 5, 10 and 50 km and the value of $\sum_i \sum_l \omega_{k,l} d_l proxli$ for each CZ (second line) for the same values of ν. All values are transformed by taking $log(1+x)$. See Section E.

Figure E2: Mapping baseline exposure



Notes: These maps report the sum of the baseline measures of exposure at the CZ level. Formally, the first map reports the value of $\sum_{l,i} d_l prox_{il}$ for each CZ (see Section 3.1. The second map does the same but replace d_l by the predicted value $\hat{d}_l$ (see Section 4.4) and the third map does the same but restricts on projects that have been rejected. All values are transformed by taking $log(1+x)$.