École Doctorale n°405
Economie, Management, Mathématiques, Physique et Sciences Informatiques

Thèse de Doctorat

pour l'obtention du titre de
Docteur en Sciences Économiques
délivré par

CY Cergy-Paris Université et l'ESSEC Business School

Connectivity and Regional Interactions: Empirical Studies on Development Disparities, Collaborative Innovation, and Mobility

présentée et soutenue publiquement le 7 décembre 2023 par

Gabrielle GAMBULI

préparée sous la direction de **Sara BIANCINI** et de **Rodrigo PAILLACAR**

Jury

Rapporteurs	
Laura HERING	Professeure associée à Erasmus University Rotterdam
Ivan LEDEZMA	Professeur à l'Université de Bourgogne
Président	
Philippe GAGNEPAIN	Professeur à Paris School of Economics
Examinatrice	
Christina TERRA	Professeure à l'ESSEC Business School
Directeurs	
Sara BIANCINI	Professeure à CY Cergy-Paris Université
Rodrigo PAILLACAR	Maître de conférence à CY Cergy-Paris Université

Connectivité et Interactions Régionales : Etudes Empiriques des Disparités de Développement, Innovation Collaborative et Mobilité

Connectivity and Regional Interactions: Empirical Studies on Development Disparities, Collaborative Innovation, and Mobility

Gabrielle GAMBULI

THEMA
CY Cergy-Paris Université
ESSEC Business School
33 boulevard du Port, 95011 Cergy, France

Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse, Sara Biancini et Rodrigo Paillacar, pour m'avoir donné la liberté et la confiance d'explorer et étudier les sujets qui m'intéressent pour cette thèse. Merci pour votre soutien et vos conseils, qui ont été déterminants dans mon parcours académique. Nos discussions et réunions ont toujours été remarquablement enrichissantes. Merci de m'avoir inspirée et encouragée tout au long du doctorat. Un clin d'œil particulier à Rodrigo, qui m'a encadrée pendant plus de cinq ans, depuis mon mémoire de master. J'ai appris tellement depuis ! Son talent d'enseignant à illustrer des mécanismes complexes avec des métaphores et des anecdotes, ainsi que sa richesse de connaissances, m'ont toujours impressionnée.

C'est un immense honneur d'avoir Philippe Gagnepain, Laura Hering, Ivan Ledezma et Christina Terra dans mon comité. Je tiens à exprimer ma plus profonde gratitude pour avoir gracieusement accepté de faire partie de cette dernière étape de mon doctorat. Un merci particulier à Laura et Ivan pour avoir généreusement accepté les rôles de rapporteurs. J'attends avec impatience de recevoir vos précieux retours, qui contribueront sans aucun doute au perfectionnement de mes recherches. Je m'excuse d'avance pour la longueur de certains chapitres. Je reconnais la nécessité de prendre le temps supplémentaire pour condenser leur contenu.

I would like to express my gratitude to my coauthor, Fernando Stipanicic, for his support and collaboration. Our mutual interest in exploring the impact of transportation on innovation resulted in insightful discussions and valuable knowledge exchange.

Du côté du Thema, j'exprime ma profonde gratitude envers Lisa, Yolande, Imen et Cécile pour avoir assuré le bon déroulement de tous les processus administratifs, rendant ainsi simples des tâches qui paraissaient insurmontables. Merci pour votre soutien et votre travail exceptionnel, grandement apprécié par l'ensemble des doctorants ! Du côté de l'ESSEC, un grand merci à Christine et Audrey, ainsi que Lina au début, pour leur écoute, leur disponibilité et leur professionnalisme exceptionnels.

Ma gratitude va également aux professeurs de Cergy, que j'ai pu avoir en cours depuis ma première année de licence. Leur influence a été déterminante dans l'éveil de ma passion pour l'enseignement et ma soif d'apprendre. Mes remerciements spéciaux vont à Pierre André, Pascal Belan, Sara Biancini, Pamela Bombarda, Olivier Donni, Fabian Gouret, Laurence Jacquet, Maëllys de la Rupelle, Rodrigo Paillacar, Nathalie Picard, Jérôme Stephan, Christina Terra, Thomas Tregouet. Un grand honneur m'a été accordé d'assurer les travaux dirigés pour Pascal, Pamela et Olivier durant mon doctorat. Ce fut un plaisir de travailler avec vous.

Merci aux (plus ou moins) 630 étudiants que j'ai eu le plaisir d'avoir en cours ces quatre dernières années. Vous avez contribué à rendre cette expérience doctorale mémorable. Une pensée particulière pour ceux que je croise encore dans les couloirs, évoluant dans leurs études et débuts de carrière. Un grand merci à ceux qui se sont particulièrement distingués en classe; voir des étudiants avides d'apprendre et interagissant avec nous est une source de satisfaction

inestimable.

Je tiens à remercier Mariona Segù et Fanny Landaud pour leur écoute attentive et leurs précieux conseils. La qualité de vos travaux et de votre présence au sein du laboratoire est une réelle source d'inspiration.

To my amazing CY-ESSEC/ESSEC-CY economic cohort — Ahmet, Huali, Margaux, Mélanie, and Thomas — thank you for having been constant companions through the exams at the beginning of the program, the challenges of the Covid pandemic, and the entire PhD journey. Mélanie et Margaux, votre soutien a été un pilier tant professionnel que personnel au cours des 5 dernières années depuis le programme de master. Mélanie, ça a été un plaisir et une douceur quotidienne de partager notre bureau. Margaux, nous avons commencé ce doctorat ensemble, et nous allons le terminer ensemble ! Merci pour ces années de blagues hilarantes, qui t'ont fait valoir le titre de doctorante la plus drôle de Cergy.

Un merci sincère à tous mes merveilleux collègues et amis du Thema. En particulier: Anderson, pour ta bienveillance et tes recommandations littéraires hors pairs. Arnaud, pour avoir été un super mentor. Ben, pour ton professionalisme en temps que collègue de TD de Macro et d'escape game. Eloïse, pour ton aide précieuse en cette fin de thèse, ainsi que pour tous les moments passés ensemble. José, pour m'avoir démontré qu'il est bénéfique de jurer en espagnol quand on a des soucis de code. Romuald, pour faire des cartes presque aussi bien que les miennes. Aurélie, pour ces sessions de tennis qui ont contribué à maintenir mon esprit sain en fin de thèse. Audin, qui a d'abord été mon tutoré, puis ami, et maintenant confrère. Avec vous, cette aventure aura été mémorable.

Un grand merci à mes chers amis : Bastien, Clara, Chloé, Lucie, Solène et Valentin (je vous mentionne par ordre alphabétique, nul besoin de spéculer sur l'ordre de mes préférences). Nos moments partagés ont été extrêmement précieux pendant ce doctorat, et depuis bien avant. Merci de booster ma santé mentale, de me faire rire autant, et d'être présents depuis toutes ces années.

Je suis profondément reconnaissante envers ma famille, qui m'a supportée pendant bien des week-end et vacances à travailler sur cette thèse. Ce doctorat est aussi en partie pour vous rendre fiers. Un merci tout particulier à mes parents. Leur intelligence, leurs réussites et leur force représentent une motivation et inspiration quotidienne. Je vous suis infiniment reconnaissante pour votre soutien et votre présence. A mon père, passionné d'histoire, qui m'a inspirée le petit passage historique dans mon introduction. A ma mère, qui m'a inspirée l'idée de travailler sur les trains à grande vitesse lors d'une conversation que nous avons eue après qu'elle ait visionné un documentaire. Maman, j'ai trois chapitres consacrés à ce sujet dans ma thèse !

À mon petit frère - Raphaël, ou Boubou pour les intimes - dont je suis immensément fière. Merci d'exister et d'être simplement toi. De plus, je reconnais qu'il n'est pas donné à tout le monde d'avoir dans sa vie quelqu'un qui sait dénicher les vidéos de chats les plus drôles d'internet.

Enfin, *last but not least*, à Lucas, mon collègue, ami, partenaire de tennis et de vie. Merci pour ton soutien indéfectible. Merci pour ton réconfort et ta patience dans les moments de doute. Merci pour la paix, la joie et les rires que tu m'apportes au quotidien.

Résumé en français

Cette thèse explore l'influence de la connectivité interrégionale sur le développement des régions et leurs interactions. La connectivité revêt diverses formes, que ce soit à travers des infrastructures physiques de transport ou la capacité d'établir des liens en raison de similitudes culturelles ou d'activités économiques partagées. Elle facilite la mobilité des personnes et des biens sur des distances variées, exerçant ainsi une influence sur les interactions et conditions économiques d'un territoire. Cette thèse se focalise sur trois dimensions spécifiques d'interactions : le commerce, les collaborations d'innovation interrégionales, et les déplacements domicile-travail des travailleurs.

Dans un premier chapitre, j'évalue l'importance de la proximité des marchés sur les différences intranationales de développement régional. Pour ce faire, je construis un indice d'accessibilité représentant le potentiel commercial de chaque région infranationale pour tous les pays du monde. Cet indice prend en compte les capacités de consommation des régions, leurs coûts de transport estimés en fonction de la proximité physique ajustée pour les contraintes géographiques, ainsi que leur proximité culturelle et économique. Une proximité accrue entre les régions a un impact positif sur leurs échanges commerciaux, qui influencent à leur tour leur niveau de développement. J'évalue également l'effet hétérogène de cette proximité sur les régions riches et périphériques.

Dans le deuxième chapitre, je m'intéresse aux répercussions d'une amélioration de la connectivité entre les régions, en explorant spécifiquement l'impact de l'introduction des lignes à grande vitesse (LGV) qui ont considérablement réduit les temps de trajet entre certaines villes. L'objectif est d'évaluer comment cette évolution influence les collaborations d'innovation entre inventeurs en France continentale. Pour ce faire, je compile une base de données qui retrace l'évolution des temps de trajet en train en France depuis 1980, avant l'implémentation de la première ligne. De plus, j'utilise une base de données de brevets français pour détecter les collaborations entre inventeurs. Les résultats indiquent que la réduction du temps de déplacement stimule simultanément le nombre, la qualité et la portée des innovations au sein de ces collaborations interrégionales.

En prolongement du deuxième chapitre, le troisième chapitre propose un exposé détaillé de la méthodologie employée pour construire les données sur l'évolution des temps de trajet en train en France de 1980 à 2017. En utilisant les horaires actuels des trains de la Société Nationale des Chemins de Fer français (SNCF), les dates d'inauguration des LGV, et l'algorithme de Dijkstra, je calcule les temps de trajet les plus courts de ville à ville en France pour chaque année. Une procédure de validation est mise en œuvre pour démontrer la fiabilité de ces données.

Dans un quatrième chapitre, j'étudie la mobilité des travailleurs en me focalisant sur les trajets domicile-travail en France continentale. L'objectif principal est d'évaluer l'impact d'une réduction du temps de trajet sur ces déplacements. Pour ce faire, j'utilise la base de données

sur les temps de trajet en train présentée dans le chapitre 3, conjointement avec les données issues du "Panel Tous Salariés" de l'INSEE pour quantifier la mobilité interrégionale des travailleurs liée à leur emploi. Mon analyse inclut également l'exploration des principaux facteurs influençant ces déplacements, ainsi que l'impact combiné de la faisabilité du télétravail. Les résultats obtenus mettent en évidence une augmentation des déplacements longue distance, tirant avantage de la réduction du temps de trajet en train. De plus, cet effet est amplifié par l'amélioration de la couverture internet dans le département de résidence. Ce constat nous offre une indice sur la complémentarité du télétravail et du réseau LGV, expliquant la croissance des déplacements longue distance.

Discipline : Sciences économiques

Classification JEL : F15, O18, O34, O36, R11, R40

Mots-clés : Géographie Economique, Intéractions Régionales, Développement Régional, Commerce International, Proximité, Connectivité, Transport, Train, Innovation, Collaboration, Mobilité, Déplacements domicile-travail, Télétravail

Summary in English

This thesis explores the influence of interregional connectivity on the development of regions and their interactions. Connectivity takes various forms, whether through physical transport infrastructure or the ability to establish connections due to cultural similarities, a shared history, and common economic activities. It facilitates the mobility of people and goods over varying distances, thus influencing economic interactions and conditions of a territory. The thesis focuses on three specific dimensions of interactions: trade, interregional innovation collaborations, and commuting patterns of workers.

In the first chapter, I assess the importance of market proximity in intranational differences in regional development. To do so, I construct an accessibility index representing the trade potential of each subnational region for all countries worldwide. This index considers the consumption capacities of regions, their transport costs estimated based on physical proximity adjusted for geographical constraints, as well as their cultural and economic proximity. Increased proximity between regions has a positive impact on their trade, which, in turn, influences their level of development. I also evaluate the heterogeneous effect of this proximity on core and peripheral regions.

In the second chapter, I delve into the repercussions of improved connectivity between regions, specifically exploring the impact of the introduction of high-speed rail (HSR) lines that have significantly reduced travel times between cities. The objective is to assess how this evolution influences innovation collaborations among inventors in mainland France. To achieve this, I compile a database tracing the evolution of train travel times in France since 1980, before the implementation of the first HSR line. Additionally, I use a database of French patents to identify collaborations between inventors. The results indicate that the reduction in travel time simultaneously stimulates the number, quality, and scope of innovations within these interregional collaborations.

As an extension of the second chapter, the third chapter provides a detailed account of the methodology used to construct data on the evolution of train travel times in France from 1980 to 2017. Using the current schedules of trains operated by the French National Railway Company (SNCF), the opening dates of HSR lines, and Dijkstra's algorithm, I calculate the shortest travel times between cities in France for each year. A validation exercise is implemented to demonstrate the reliability of this data.

In the fourth chapter, I study the mobility of workers by focusing on commuting patterns in mainland France. The main objective is to assess the impact of reduced travel time on these commutes. To achieve this, I use the database on train travel times presented in Chapter 3, along with data from the "Panel Tous Salariés" of INSEE to quantify the interregional mobility of workers related to their employment. My analysis also includes exploring the major factors influencing these commutes, as well as the combined impact of the feasibility of telecommuting. The results highlight an increase in long-distance commutes, taking advantage of the

reduced train travel time. Furthermore, this effect is amplified by the improvement of internet coverage in the residential department. This observation provides an indication of the complementarity of telecommuting and the HSR network, explaining the growth of long-distance commutes.

Field: Economics

JEL Classification: F15, O18, O34, O36, R11, R40

Keywords: Economic Geography, Regional Interactions, Regional Development, International Trade, Proximity, Connectivity, Transportation, Train, Innovation, Collaboration, Mobility, Commuting, Teleworking

Contents

List of Tables

List of Figures

Introduction Générale

La **distance** géographique a indéniablement exercé une influence prépondérante tout au long de l'histoire de l'humanité, jouant un rôle essentiel dans les déplacements des individus, notre appréhension du monde, la propagation et le développement des cultures, ainsi que dans les dynamiques d'échanges commerciaux.

Alors que la distance exprime la proximité ou l'éloignement physique entre deux points géographiques, les avancées technologiques dans les domaines du transport et de la communication nous conduisent plutôt à adopter le terme de **connectivité** dans l'étude des interactions et échanges économiques. La connectivité exprime la capacité à relier ou interconnecter des éléments. Elle peut dépendre de la distance, via les transports, ou en être totalement exempt, ou presque, grâce aux technologies de communication comme internet.

Cette thèse explore comment la connectivité, en facilitant la mobilité des personnes et des biens sur diverses distances, influence le développement territorial et leurs interactions. Elle se concentre notamment sur trois dimensions d'interactions spécifiques : **le commerce**, sujet du premier chapitre, **les collaborations d'innovation**, explorées dans le deuxième chapitre, et **les déplacements domicile-travail des travailleurs**, traités dans le quatrième chapitre.

Le troisième chapitre présente une note sur la création d'une nouvelle base de données suivant l'évolution de la connectivité ferroviaire des villes et régions françaises, en lien avec l'implémentation d'une nouvelle technologie de transport: les lignes et trains à grande vitesse. Ces données sont utilisées pour les études du deuxième et quatrième chapitre.

Un peu d'histoire.[1] Au cours des premières phases de l'histoire humaine, les migrations massives, amorcées il y a 2 millions d'années, ont pris siècles et millénaires à s'étendre géographiquement, entravées par la lenteur des déplacements à pied. Il a fallu attendre 800 000 années pour voir les premiers hommes atteindre l'Europe, et le début de sa colonisation il y a 50 000 ans.

La sédentarisation ultérieure, entre 10 000 et 2 000 avant JC, a restreint les déplacements des individus à des distances relativement courtes. Cette époque a donné naissance aux premières routes et à l'invention de la roue. Ces avancées rudimentaires ont facilité les déplacements locaux entre villages et lieux d'élevage. Les échanges se limitaient aux trajets terrestres, mais les fondements des futurs réseaux de connectivité étaient posés.

Les progrès technologiques ultérieurs ont ouvert la voie à des voyages et explorations plus étendues. Les Phéniciens, grands navigateurs de l'Antiquité, ont établi des routes maritimes à courte distance en Méditerranée, créant ainsi des liens commerciaux dans cette région, participant également à la diffusion de l'écriture et du langage.

L'émergence des premiers empires sur de vastes distances géographiques, tels que l'Empire

[1] Source: *L'histoire du monde par les cartes*, Larousse, édition 2020.

de Chine sous la dynastie Qin, l'Empire Romain, et l'Empire Byzantin, a été grandement facilitée par les avancées technologiques dans le domaine du transport. Un exemple emblématique de cette époque est la conception des voies romaines, permettant de relier des territoires distants, facilitant les échanges ainsi que la cohésion et gouvernance de l'Europe. Au Moyen Âge, avec la Route de la Soie les distances parcourues ont augmenté, facilitant les échanges entre l'Orient et l'Occident.

Au XV^{ème} et XVI^{ème} siècle, c'est l'essor du transport maritime et des grands navigateurs européens. Ces explorateurs, bénéficiant de navires robustes et imposants, ont ouvert de nouvelles routes maritimes vers les autres continents. Cette expansion a considérablement élargi les échanges à l'échelle mondiale, intensifiant les liens entre les continents, notamment à travers la mise en place d'empires coloniaux.

L'ère des échanges, du XVI^{ème} au XIX^{ème} siècle, a été marquée par des flux croissants, stimulés par la révolution industrielle. L'invention du moteur à vapeur, le développement des constructions navales et des chemins de fer ont intensifié les échanges nationaux et mondiaux. De plus, l'invention du télégraphe a considérablement réduit les délais de transmission d'informations sur des milliers de kilomètres.

Cette période a également transformé la géographie humaine infranationale. Les usines, libérées de la nécessité d'être proches des matières premières nécessaires à la production, se sont installées en ville, provoquant un exode rural massif et une augmentation considérable de la population urbaine. Les transports urbains, dont le premier métro à Londres, ont émergé pour répondre aux besoins croissants des travailleurs.

La mondialisation moderne a pris de l'ampleur à partir de la seconde moitié du XX^{ème} siècle. Les transports mondiaux, tels que les porte-conteneurs, les avions-cargos, les réseaux routiers et ferroviaires interconnectés ont facilité la libre circulation des biens, des personnes, de l'argent, des connaissances et de la culture à travers le monde.

Cette avancée des technologies de transport, ainsi que l'émergence des technologies d'information et de communication, notamment internet, ont profondément transformé la dynamique des interactions humaines, réduisant considérablement les contraintes traditionnelles liées aux barrières physiques de l'espace. Ce phénomène est illustré par le concept de *mort de la distance* proposé par Cairncross (1997), qui appréhende cette nouvelle capacité des individus de se déplacer, commercer, communiquer, et collaborer sur de vastes distances.

Et du côté de la littérature économique? La première conceptualisation d'un phénomène économique lié à la distance est souvent attribuée à Hotelling (1929), dans un cadre de concurrence. En illustrant un modèle de localisation d'entreprise le long d'un segment, il démontre que les entreprises, qui se positionnent stratégiquement pour maximiser leur part de marché en minimisant la distance moyenne aux consommateurs, finissent par se retrouver toutes les deux au centre du segment. D'un autre côté, Weber and Friedrich (1929) explique que les entreprises choisissent des emplacements qui minimisent les coûts de production, en prenant en compte des facteurs tels que les coûts de transport, de main-d'œuvre et d'autres coûts logistiques. A la même époque, entre 1980 and 1920, Marshall (1890) explique la tendance des entreprises à se regrouper dans des zones géographiques spécifiques qui favorise l'échange d'idées, de travailleurs qualifiés et de technologies.

Ces modèles ont introduit les bases du concept de *grappe industrielle*, ou *industrial cluster* en anglais, qui souligne les avantages de la proximité géographique des entreprises similaires, tels que l'amélioration de la productivité résultant de la concurrence et de l'accès facilité aux

savoirs, favorisant ainsi la création d'économies d'échelle. C'est avec la *Nouvelle Géographie Economique*, initiée par Krugman (1980, 1991), que l'on commence à regarder au-delà des frontières nationales pour comprendre comment les économies d'échelle, les coûts de transport et les barrières commerciales forment le commerce international.

Le travail de Krugman a souligné l'impact significatif de l'accès aux marchés, c'est-à-dire l'accès à la demande (aux consommateurs), sur la répartition spatiale de l'activité économique. C'est dans ce contexte que Fujita et al. (1999) ont développé un modèle d'équilibre général qui explique les inégalités de revenus par la localisation des entreprises, en supposant une main d'œuvre immobile. Dans le cadre d'une concurrence monopolistique, les entreprises produisent des biens différenciés et opèrent sous des rendements d'échelle croissants, qui les incitent à produire de façon conséquente pour exporter leurs bien à l'internationale pour ainsi générer plus de profits. Plus une entreprise est située près de la demande nationale et internationale, plus elle réalise de profits, et mieux sa main-d'œuvre est rémunérée.

Ainsi, le modèle de Fujita et al. (1999) expose une corrélation positive entre les revenus des facteurs et l'accès au marché, également connue sous le nom d'*équation des salaires du commerce international* ou *international trade wage equation* en anglais. Cette relation a été explorée au niveau national (Redding and Venables, 2004; Head and Mayer, 2011), expliquant les différences de développement entre pays, et a également reçu une attention au niveau infranational, pour expliquer les différences de développement au sein même des pays ou en Europe (Brakman et al., 2004; Hering and Poncet, 2010; Brakman et al., 2009). Le **premier chapitre** de cette thèse élargit cette enquête au niveau infranational en utilisant des données régionales couvrant l'ensemble des pays du monde, afin de fournir un test de falsification global.

Au tournant du millénaire, l'économie géographique a intégré la mobilité des facteurs de production, mettant notamment en avant les migrations de main-d'œuvre. Dans ce contexte, Koser (2007) analyse les processus, causes et conséquences de la migration internationale, en explorant les motivations telles que les opportunités économiques, l'instabilité politique et le désir d'une meilleure qualité de vie. Il examine l'impact de la migration sur les pays d'origine et de destination, influençant la dynamique économique, les marchés du travail et les structures sociales. D'autre part, Borjas (2014) se penche sur l'impact de l'afflux de travailleurs immigrés sur les salaires, les opportunités d'emploi et les résultats professionnels des travailleurs locaux et d'autres immigrants. Il aborde la segmentation du marché du travail et explore le rôle des niveaux d'éducation et de compétences dans les conséquences économiques de l'immigration.

Ces questions sont également pertinentes dans le cadre de la migration infranationale, bien moins coûteuse que la migration internationale, et donc moins rare. Le cas des États-Unis a suscité une attention particulière dans les études (Chiquiar and Hanson, 2005; Kennan and Walker, 2011; Monte et al., 2018), dû à la flexibilité remarquable des travailleurs américains dans la capacité à saisir de nouvelles opportunités malgré les distances à parcourir. En Europe, ce phénomène est moins manifeste en raison de la diversité culturelle et linguistique, même au sein d'un espace de libre circulation des biens et des personnes (Faist, 2000; Van Houtum and Van Der Velde, 2004). Les coûts liés à la distance physique et culturelle semblent ainsi demeurer significatifs, contrairement aux anticipations de Cairncross (1997).

Alors que la plupart des recherches ont porté sur la migration impliquant le déplacement du lieu de vie et de travail d'un lieu A à un lieu B, en fonction du salaire et du coût du logement, un autre type de migration, jusqu'ici peu étudié, est possible. Il s'agit d'une migration où les individus choisissent de vivre à un lieu B et de travailler à un lieu C, modifiant soit un seul,

soit les deux composants de leur situation initiale. Cette migration, qu'elle soit partielle ou complète, devient envisageable grâce à la possibilité d'effectuer des trajets domicile-travail, appelé *commuting*, sur de longues distances.

Les études sur le *commuting* s'appuient principalement sur des modèles de structure urbaine monocentrique ou métropolitaine, où les individus travaillent majoritairement au centre, concentrant les emplois principaux, et résident dans des zones s'étendant de manière décroissante vers la périphérie, où les logements sont plus abordables (Alonso, 1964; de Palma et al., 2007). Cependant, l'émergence d'une nouvelle forme de mobilité longue-distance entre centres urbains éloignés est rendue possible et observée grâce à des infrastructures de transport particulièrement rapides, les lignes et trains à grande vitesse (Guirao et al., 2018; Heuermann and Schmieder, 2019; Wang et al., 2019).[2] Ces avancées technologiques peuvent profondément influencer la répartition des résidents et des travailleurs sur le territoire. C'est précisément ce qu'examine le **quatrième chapitre** de cette thèse pour le cas de la France.

Un autre pan de la littérature se focalise sur l'innovation et son rôle dans le développement régional. Les chercheurs prennent inspiration de Marshall (1890), Jacobs (1969) et Porter (1990), mettant en lumière comment la proximité spatiale d'entreprises peut catalyser l'innovation (Boschma, 2005). Les *clusters* d'innovation sont devenus des épicentres d'une croissance économique dynamique, comme indiqué par le modèle de Akcigit et al. (2018), jetant une lumière nouvelle sur la manière dont la créativité et l'entrepreneuriat peuvent transformer les régions.

Cependant, bien que les clusters offrent de nombreux avantages en termes de collaboration et d'échange de connaissances, ils peuvent également conduire à un effet de verrouillage, appelé *lock-in effect*, comme souligné par Boschma (2005). Cet effet suggère qu'après une certaine période de collaboration étroite, il devient de plus en plus difficile pour les individus et les entreprises d'acquérir de nouvelles connaissances et de réaliser des progrès significatifs sur des projets innovants. Ce phénomène peut limiter la diversité des idées et des perspectives, entravant potentiellement l'innovation et la croissance.

La connectivité par les transports, qu'ils soient routiers, ferroviaires ou encore aériens, représente une voie cruciale pour maintenir l'ouverture aux sources externes de connaissances. Des études ont examiné comment cette connectivité influe sur l'innovation, les citations et les collaborations (Bernard et al., 2020; Catalini et al., 2020; Koh et al., 2022; Pauly and Stipanicic, 2022; Tsiachtsiras, 2022; Andersson et al., 2023). Une contribution récente à cette littérature émerge de la Chine, où l'effet des lignes à grande vitesse sur la capacité à connecter des inventeurs de régions différentes est exploré (Hanley et al., 2022; Li et al., 2022; Yao and Li, 2022; Kang et al., 2023). Le **deuxième chapitre** de cette thèse se penche sur cette dynamique, en se concentrant sur le cas de la France.

Les études antérieures sur l'impact des trains à grande vitesse en Allemagne et en Chine reposent sur des bases de données existantes (Heuermann and Schmieder, 2019; Wang et al., 2019; Kang et al., 2023), ou utilisent simplement des variables indicatrices signalant la présence d'une ligne ou d'une station les reliant, comme dans l'étude de Guirao et al. (2018) en Espagne. Mes recherches sur le contexte français, explorées dans les deuxième et quatrième chapitres de cette thèse, sont rendues possibles par l'élaboration d'une base de données inédite retraçant l'évolution des temps de trajet en train entre les villes françaises depuis 1980, année avant la construction de la première ligne à grande vitesse. Le **troisième chapitre** de cette thèse est dédié à la présentation de cette base de données.

Ces dernières années ont été marquées par l'intrusion croissante des technologies de l'information

[2] Les références mentionnées ici sont des études de cas menées en Espagne, en Allemagne et en Chine.

dans notre quotidien, inaugurant un nouveau chapitre dans notre compréhension de la manière dont la connectivité numérique influence non seulement les échanges et les flux d'information, mais également la structure spatiale de notre société. Cet impact est amplifié par l'émergence et l'adoption généralisée des pratiques de télétravail, déjà présentes avant la pandémie de COVID-19, mais considérablement intensifiées par la suite. Cette thèse intègre ces dimensions nouvelles dans les deuxième et quatrième chapitres.

Cette thèse vise à tirer parti des données de plus en plus détaillées en les associant à un savoir-faire informatique, dans le but de développer des indices et des mesures originales qui servent à enrichir les analyses économétriques existantes. Ci-dessous sont présentés brièvement chacun de mes chapitres, leur méthodologie et résultats principaux.

Le **premier chapitre** de cette thèse examine l'impact de la proximité des marchés sur le développement infranational à l'échelle mondiale, en tenant compte des effets hétérogènes sur les régions centrales et périphériques, ainsi que sur les pays présentant différents niveaux de revenus. Un nouvel indice de potentiel de marché basé sur la gravité est proposé pour évaluer avec précision les distances terrestres et maritimes afin de mieux saisir les limites géographiques dans notre confrontation à la distance. Les estimations sont réalisées en coupe transversale avec des effets fixes au niveau des pays, en traitant les problèmes d'endogénéité avec des variables instrumentales et variables de substitution. Des vérifications de robustesse sont également effectuées avec des données de panel sur un échantillon plus restreint.

Les résultats révèlent que les régions bénéficiant d'un meilleur accès aux marchés et d'un bon accès portuaire enregistrent un revenu régional par habitant plus élevé, l'effet étant plus prononcé pour les régions plus riches. Les régions périphériques présentent une élasticité de 2 points de pourcentage inférieure au potentiel de marché par rapport aux régions centrales. Le chapitre souligne également l'impact négatif potentiel de la proximité des marchés étrangers sur les régions périphériques, surtout si elles sont proches des marchés étrangers centraux sans accord de libre-échange entre les pays respectifs. Les résultats suggèrent que les politiques visant à améliorer la connectivité des régions périphériques aux marchés domestiques centraux et à développer des accords commerciaux pourraient contribuer à atténuer les effets néfastes des barrières commerciales et à réduire les disparités régionales au sein des pays.

Le **deuxième chapitre** de cette thèse, en collaboration avec Fernando Stipanicic, explore l'impact des trains à grande vitesse (TGV) sur les collaborations entre inventeurs, jouant le rôle de catalyseur pour les interactions en face-à-face et l'échange de connaissances sur de longues distances. Nous utilisons un nouvel ensemble de données sur les temps de trajet suite à la mise en place du TGV en France. En utilisant un modèle gravitationnel avec des effets fixes tridirectionnels, nous évaluons la relation causale entre la réduction du temps de trajet et les tendances de collaboration pour brevets entre paires de départements, en traitant les préoccupations d'endogénéité.

Les résultats montrent un effet positif robuste de la réduction du temps de trajet sur les collaborations inter-départementales. Une meilleure connectivité sur de longues distances favorise la création de nouvelles collaborations et renforce celles déjà en place. Une analyse de l'effet hétérogène de l'effet de la réduction du temps de trajet révèlent que seulement les régions centrales voient leur nombre de collaborations augmenter, mais les avantages s'étendent au-delà des régions directement connectées par le TGV. De plus, la réduction du temps de trajet

est associée à des brevets collaboratifs qui se distinguent par leur nouveauté et leur diversité technologique accrues. Enfin, il est à noter que la diminution du temps de trajet a facilité les connexions entre tous types d'inventeurs, avec une intensité plus marquée pour ceux ayant une productivité supérieure à la moyenne de leur localité respective.

Le **troisème chapitre** de cette thèse, en collaboration avec Fernando Stipanicic, présente un nouvel ensemble de données sur les temps de trajet en train en France, couvrant la période de 1980 à 2020, ainsi que la méthodologie suivie pour le construire. Nous avons utilisé un emploi du temps sur les arrivées et départs de trains de la Société Nationale des Chemins de Fer (SNCF), ainsi que les dates d'ouverture des lignes à grande vitesse (LGV), construites de 1981 à 2017 pour relier Paris aux principales villes de France.

En utilisant l'algorithme de Dijkstra (Dijkstra, 1959), nous calculons le temps de trajet contemporain entre chaque paire de villes en France. Ensuite, afin de calculer les temps de trajet historiques au sein de chaque paire, nous partons de l'hypothèse que la circulation des trains s'effectue à vitesse normale avant l'ouverture d'une LGV. Nous sommes en mesure de comparer nos estimations du temps de trajet en train aux valeurs observées du temps de trajet à partir d'un sous-échantillon de paires de villes (SNCF). Nous trouvons que notre base de données reproduit 95% de la variation totale du temps de trajet observé.

Enfin, le **quatrième chapitre** de cette thèse étudie l'impact du réseau ferroviaire à grande vitesse en France sur les décisions de déplacement et de déménagement des travailleurs concernant leur lieu de travail et leur résidence. Ce travail utilise des données sur les travailleurs français issues du Panel DADS Tout Salarié, ainsi que des données sur le temps de trajet en train présenté dans le troisième chapitre.

Je quantifie l'impact d'une réduction du temps de trajet sur les ajustements des schémas de déplacement domicile-travail entre départements français en utilisant un modèle gravitaire. Afin de prendre en compte le fait que le temps de trajet demeure important sur les paires de départements affectées par les LGV, j'intègre dans le modèle la dimension du télétravail et la connectivité à Internet. J'aborde cette question en segmentant l'analyse par catégorie socio-professionnelle des travailleurs, étant donné que ces groupes n'ont pas tous la même possibilité de télétravail. Cette analyse reconnaît le rôle essentiel d'Internet et du télétravail dans la réduction des coûts liés aux déplacements quotidiens.

Les résultats montrent une tendance croissante à la localisation du domicile et du lieu de travail dans des départements distincts et distants, pour toute catégorie de travailleurs, que ce soit les cadres, les employés et les ouvriers. Cette tendance provient à la fois d'une meilleure connectivité des territoires par les LGV, mais également d'une meilleure connectivité à internet, qui joue à la fois tel une commodité souhaitable pour les individus, ainsi qu'un complément des LGV pour réduire les coûts associés aux déplacements domicile-travail, sous l'hypothèse qu'un accès internet rend le télétravail possible.

Une version future de cette analyse examinera en détail les facteurs additionnels qui contribuent à ces changements de schémas. Cela inclura (1) une mise en évidence particulière sur les types de transitions, que ce soit des environnements urbains vers ruraux ou vice versa, (2) une exploration de la flexibilité des variables d'ajustement, que ce soit le lieu de résidence, le lieu de travail, ou les deux simultanément; (3) une analyse de la mobilité des travailleurs au sein d'une même entreprise présente à plusieurs endroits ou entre différentes entreprises, et enfin (4) une évaluation de la différence de salaires associée au changement de lieu de travail.

Conclusion Générale. Cette thèse, inscrite dans la continuité de la recherche existante tout en visant à contribuer à son enrichissement, a démontré que l'atténuation de l'effet de la distance au fil du temps est attribuable à une amélioration de la connectivité, facilitée par la mise en œuvre d'accords de libre-échange entre les nations ou par l'évolution des technologies telles que les lignes à grande vitesse et internet. Cependant, il convient de noter que la connectivité des territoires est inégale et que les barrières inhérentes à la distance conservent leur impact, comme en témoignent la concentration de l'activité économique sur certains territoires et la persistance des régions périphériques. En somme, la distance demeure une variable significative, bien que la progression humaine vers des horizons toujours plus lointains demeure une réalité constante.

General Introduction

The geographical **distance** has undoubtedly exerted a significant influence throughout human history, playing a crucial role in individuals' movements, our perception of the world, the spread and development of cultures, as well as in the dynamics of trade.

While distance expresses the physical proximity or remoteness between two geographical points, technological advancements in the fields of transportation and communication lead us to prefer the term **connectivity** in the study of economic interactions and exchanges. Connectivity expresses the ability to link or interconnect elements. It may depend on distance, through transportation, or be almost entirely independent, thanks to communication technologies like the internet.

This thesis explores how connectivity, by facilitating the mobility of people and goods over various distances, influences territorial development and its interactions. It focuses specifically on three dimensions of interactions: **trade**, the subject of the first chapter, **innovation collaborations**, explored in the second chapter, and **commuting patterns of workers**, addressed in the fourth chapter.

The third chapter presents a note on the creation of a new database tracking the evolution of railway connectivity between French cities and regions, following with the implementation of a new transportation technology: high-speed lines and trains. These data are used for the studies in the second and fourth chapters.

A bit of history.[3] During the early phases of human history, massive migrations, initiated around 2 million years ago, took centuries and millennia to spread geographically, hindered by the slow pace of travel on foot. It took 800,000 years for the first humans to reach Europe, marking the beginning of its colonization 50,000 years ago.

Subsequent sedentarization, between 10,000 and 2,000 BC, restricted individual movements to relatively short distances. This era gave rise to the first roads and the invention of the wheel. These rudimentary advances facilitated local travel between villages and pastoral locations. Exchanges were limited to terrestrial routes, but the foundations of future connectivity networks were laid.

Later technological progress paved the way for more extensive travel and exploration. The Phoenicians, great navigators of antiquity, established short-distance maritime routes in the Mediterranean, creating trade links in the region and contributing to the spread of writing and language.

The emergence of the first empires over vast geographical distances, such as the Chinese Empire under the Qin Dynasty, the Roman Empire, and the Byzantine Empire, was greatly facilitated by technological advances in transportation. An emblematic example of this pe-

[3] Source: *L'histoire du monde par les cartes*, Larousse, 2020 edition.

riod is the design of Roman roads, connecting distant territories, facilitating exchanges, and promoting the cohesion and governance of Europe. In the Middle Ages, with the Silk Road, traveled distances increased, facilitating exchanges between the East and the West.

In the 15th and 16th centuries, there was the rise of maritime transport and European explorers. These explorers, with robust and imposing ships, opened new maritime routes to other continents. This expansion significantly expanded global exchanges, intensifying connections between continents, especially through the establishment of colonial empires.

The era of exchanges, from the 16th to the 19th century, was marked by increasing flows, spurred by the Industrial Revolution. The invention of the steam engine, the development of shipbuilding, and railways intensified national and global exchanges. Moreover, the invention of the telegraph significantly reduced information transmission times over thousands of kilometers.

This period also transformed subnational human geography. Factories, freed from the need to be close to the raw materials necessary for production, settled in cities, causing a massive rural exodus and a considerable increase in the urban population. Urban transportation, including the first subway in London, emerged to meet the growing needs of workers.

Modern globalization gained momentum from the second half of the 20th century. Global transportation, such as container ships, cargo planes, interconnected road and rail networks, facilitated the free movement of goods, people, money, knowledge, and culture worldwide.

This advance in transportation technologies, along with the emergence of information and communication technologies, notably the Internet, has profoundly transformed the dynamics of human interactions, significantly reducing traditional constraints related to physical barriers of space. This phenomenon is illustrated by the concept of the *death of distance* proposed by Cairncross (1997), which captures this new ability of individuals to travel, trade, communicate, and collaborate over vast distances.

And from the economic literature perspective? The first conceptualization of an economic phenomenon related to distance is often attributed to Hotelling (1929) in a competitive framework. Illustrating a model of firm location along a segment, he demonstrates that companies strategically position themselves to maximize their market share by minimizing the average distance to consumers, ultimately leading both companies to end up at the center of the segment. On the other hand, Weber and Friedrich (1929) explains that companies choose locations that minimize production costs, taking into account factors such as transportation costs, labor costs, and other logistical costs. During the same period, between 1980 and 1920, Marshall (1890) explains the trend of companies to cluster in specific geographical areas due to their ability to foster the exchange of ideas, skilled workers, and technologies.

These models introduced the foundation of the *industrial cluster* concept, emphasizing the benefits of the geographical proximity of similar businesses, such as improved productivity resulting from competition and easier access to knowledge, thus fostering economies of scale. It was with the *New Economic Geography*, initiated by Krugman (1980, 1991), that researchers began to look beyond national borders to understand how economies of scale, transportation costs, and trade barriers shape international trade.

Krugman's work highlighted the significant impact of market access, i.e., access to demand (consumers), on the spatial distribution of economic activity. In this context, Fujita et al. (1999) developed their general equilibrium model explaining income inequalities based on the location of businesses, assuming immobile labor. In a monopolistic competition framework,

companies produce differentiated goods and operate under increasing returns to scale, encouraging them to produce consistently to export their goods internationally and generate more profits. The closer a company is to national and international demand, the more profits it realizes, and the better its workforce is remunerated.

Fujita et al. (1999)'s model reveals a positive correlation between factor incomes and market access, also known as the *international trade wage equation*. This relationship has been explored at the national level (Redding and Venables, 2004; Head and Mayer, 2011), explaining development differences between countries, and has also received attention at the subnational level to explain development differences within countries or in Europe (Brakman et al., 2004; Hering and Poncet, 2010; Brakman et al., 2009). The **first chapter** of this thesis extends this investigation to the subnational level using regional data covering all countries worldwide to provide a comprehensive falsification test.

At the turn of the millennium, economic geography incorporated the mobility of factors of production, notably highlighting labor migrations. In this context, Koser (2007) analyzes the processes, causes, and consequences of international migration, exploring motivations such as economic opportunities, political instability, and the desire for a better quality of life. He examines the impact of migration on both origin and destination countries, influencing economic dynamics, labor markets, and social structures. On the other hand, Borjas (2014) delves into the impact of the influx of immigrant workers on wages, employment opportunities, and professional outcomes for local workers and other immigrants. He addresses labor market segmentation and explores the role of education levels and skills in the economic consequences of immigration.

These questions are also relevant in the context of subnational migration, which is less costly than international migration and therefore less rare. The case of the United States has attracted particular attention in studies (Chiquiar and Hanson, 2005; Kennan and Walker, 2011; Monte et al., 2018), owing to the remarkable flexibility of American workers in seizing new opportunities despite the distances involved. In Europe, this phenomenon is less evident due to cultural and linguistic diversity, even within a space of free movement of goods and people (Faist, 2000; Van Houtum and Van Der Velde, 2004). The costs associated with physical and cultural distance seem to remain significant, contrary to the expectations of Cairncross (1997).

While most research has focused on migration involving the relocation of both residence and workplace from place A to place B, based on salary and housing costs, another type of migration, until now less studied, is possible. This involves individuals choosing to live in place B and work in place C, modifying either one or both components of their initial situation. This migration, whether partial or complete, becomes feasible due to the possibility of commuting over long distances.

Studies on commuting primarily rely on models of monocentric or metropolitan urban structure, where individuals predominantly work in the center, concentrating the main jobs, and reside in areas extending decreasingly towards the periphery, where housing is more affordable (Alonso, 1964; de Palma et al., 2007). However, the emergence of a new form of long-distance mobility between distant urban centers is made possible and observed through particularly fast transport infrastructure, namely high-speed rail lines and trains (Guirao et al., 2018; Heuermann and Schmieder, 2019; Wang et al., 2019).[4] These technological advances can profoundly influence the distribution of residents and workers across the territory. This is precisely what the **fourth chapter** of this thesis examines for the case of France.

[4]The references mentioned here are case studies conducted in Spain, Germany, and China.

Another aspect of the literature on Economic Geography focuses on innovation and its role in regional development. Researchers draw inspiration from Marshall (1890), Jacobs (1969), and Porter (1990), highlighting how the spatial proximity of businesses can catalyze innovation (Boschma, 2005). Innovation clusters have become epicenters of dynamic economic growth, as indicated by the model of Akcigit et al. (2018), shedding new light on how creativity and entrepreneurship can transform regions.

However, although clusters offer many benefits in terms of collaboration and knowledge exchange, they can also lead to a lock-in effect, as emphasized by Boschma (2005). This effect suggests that after a certain period of close collaboration, it becomes increasingly difficult for individuals and companies to acquire new knowledge and make significant progress on innovative projects. This phenomenon can limit the diversity of ideas and perspectives, potentially hindering innovation and growth.

Connectivity through transportation, whether by road, rail, or air, represents a crucial path to maintaining openness to external sources of knowledge. Studies have examined how this connectivity influences innovation, citations, and collaborations (Bernard et al., 2020; Catalini et al., 2020; Koh et al., 2022; Pauly and Stipanicic, 2022; Tsiachtsiras, 2022; Andersson et al., 2023). A recent contribution to this literature emerges from China, where the effect of high-speed rail lines on the ability to connect inventors from different regions is explored (Hanley et al., 2022; Li et al., 2022; Yao and Li, 2022; Kang et al., 2023). The **second chapter** of this thesis delves into this dynamic, focusing on the case of France.

Previous studies on the impact of high-speed trains in Germany and China rely on existing databases (Heuermann and Schmieder, 2019; Wang et al., 2019; Kang et al., 2023), or simply use indicator variables signaling the presence of a line or station connecting them, as in the study by Guirao et al. (2018) in Spain. My research on the French context, explored in the second and fourth chapters of this thesis, is made possible by the development of a unique database tracing the evolution of train travel times between French cities since 1980, a year before the construction of the first line. The **third chapter** of this thesis is dedicated to presenting this database.

In recent years, there has been an increasing intrusion of information technologies into our daily lives, ushering in a new chapter in our understanding of how digital connectivity influences not only exchanges and information flows but also the spatial structure of our society. This impact is amplified by the emergence and widespread adoption of telecommuting practices, already present before the COVID-19 pandemic but significantly intensified thereafter. This thesis incorporates these new dimensions in the second and fourth chapters.

This thesis aims to leverage increasingly detailed data by combining it with computational expertise, with the goal of developing original indices and measures that contribute to enriching existing econometric analyses. Below, each of my chapters is briefly presented, along with their methodology and main results.

The **first chapter** of this thesis examines the impact of market proximity on subnational development globally, considering heterogeneous effects on central and peripheral regions, as well as countries with different income levels. A new market potential index based on gravity is proposed to accurately assess terrestrial and maritime distances, better capturing geographical boundaries in our confrontation with distance. The estimates are conducted in cross-section with country-level fixed effects, addressing endogeneity issues with instrumental variables

and substitution variables. Robustness checks are also performed with panel data on a more restricted sample.

The results reveal that regions with better access to markets and good port access record higher regional income per capita, with the effect being more pronounced for wealthier regions. Peripheral regions consistently exhibit a 2 percentage point lower elasticity to market potential compared to central regions. The chapter also highlights the potential negative impact of proximity to foreign markets on peripheral regions, especially if they are close to central foreign markets without a free trade agreement between the respective countries. The results suggest that policies aimed at improving connectivity of peripheral regions to central domestic markets and developing trade agreements could help mitigate the adverse effects of trade barriers and reduce regional disparities within countries.

The **second chapter** of this thesis, in collaboration with Fernando Stipanicic, explores the impact of high-speed railways (HSR) on collaborations between inventors, acting as a catalyst for in-person interactions and knowledge exchange over long distances. We use a new dataset on travel times following the introduction of the HSR in France. Using a gravity model with tri-directional fixed effects, we assess the causal relationship between reduced travel time and trends in patent collaboration between pairs of departments, addressing endogeneity concerns.

The results show a robust positive effect of reduced travel time on inter-departmental collaborations. Better long-distance connectivity promotes the creation of new collaborations and strengthens existing ones. Central regions benefit significantly, but the advantages extend beyond regions directly connected by HSR. Additionally, reduced travel time is associated with collaborative patents distinguished by increased novelty and technological diversity. Finally, it is noteworthy that the decrease in travel time facilitated connections between all types of inventors, with a more pronounced intensity for those with above-average productivity in their respective locality.

The **third chapter** of this thesis, in collaboration with Fernando Stipanicic, presents a new dataset on train travel times in France, covering the period from 1980 to 2020, as well as the methodology followed to construct it. We used a schedule of arrivals and departures of trains from the National Railway Company (SNCF), as well as the opening dates of high-speed railways (HSR), built from 1981 to 2017 to connect Paris to major cities in France.

Using Dijkstra's algorithm (Dijkstra, 1959), we calculate contemporary travel times between each pair of cities in France. Then, to calculate past values of travel time within each pair, we rely on the assumption of train circulation at a normal speed before the opening of an LGV. We are able to compare our estimates of train travel time to observed values from a subsample of city pairs (SNCF). We find that our database reproduces 95% of the total observed variation in travel time.

Finally, the **fourth chapter** of this thesis examines the impact of the high-speed rail network in France on the travel and relocation decisions of workers regarding their workplace and residence. This work uses data on French workers from the DADS Panel Tout Salarié, as well as data on train travel time presented in the third chapter.

I quantify the impact of a reduction in travel time on adjustments to home-to-work commuting patterns between French departments using a gravity model. To account for the fact that travel time remains significant on department pairs affected by the HSR, I integrate into

the model the dimension of telecommuting and internet connectivity. I address this issue by segmenting the analysis by socio-professional category of workers, as these groups do not all have the same telecommuting possibilities. This analysis recognizes the essential role of the Internet and telecommuting in reducing costs associated with daily commuting.

The results show an increasing trend in locating homes and workplaces in separate and distant departments for all categories of workers, including executives, employees, and blue-collar workers. This trend arises from both improved territorial connectivity through HSR and enhanced internet connectivity. Internet is found to serve as both a desirable convenience for individuals and a complement to high-speed railways, reducing costs associated with home-to-work commuting, assuming that internet access makes telecommuting possible.

A future version of this analysis will delve into additional factors contributing to these pattern changes. This will include (1) a particular focus on types of transitions, whether from urban to rural environments or vice versa, (2) an exploration of the flexibility of adjustment variables, whether it be place of residence, place of work, or both simultaneously; (3) an analysis of worker mobility within the same company located in multiple places or between different companies, and finally (4) an assessment of the wage difference associated with a change in workplace, comparing levels at the destination to the origin.

General Conclusion. This thesis, building upon existing research while aiming to contribute to its enrichment, has demonstrated that the attenuation of the distance effect over time is attributable to improved connectivity, facilitated by the implementation of free-trade agreements between nations or the advancement of technologies such as high-speed rail and the internet. However, it is noteworthy that the connectivity of territories is uneven, and the inherent barriers of distance retain their impact, evident in the concentration of economic activity in certain areas and the persistence of peripheral regions. To put it in a nutshell, distance remains a significant variable, although human progress towards ever more distant horizons remains a constant reality.

Chapter 1

Navigating the Geography of Regional Disparities: Market Access and the Core-Periphery Divide

Abstract

This paper investigates the impact of market proximity on subnational development worldwide, considering the heterogeneous effects on core and peripheral regions, as well as on countries with different income levels. A gravity-based market potential index is revised to accurately assess distances for land and maritime trips to better capture geographic limitations. Estimations are performed in cross-section with country-fixed effects, by addressing endogeneity issues with instrumental variables. Robustness checks are also conducted with panel data on a smaller sample. The findings reveal that regions with better access to markets and port experience higher regional income per capita, with the effect being higher for wealthier regions. Peripheral regions consistently exhibit a 2 percentage point lower elasticity to market potential compared to core regions. The paper also highlights the potential negative impact of proximity to foreign markets on peripheral regions, especially if they are central to foreign core markets without free trade agreement in place between the respective countries. Results suggest that policies which aim at improving the connectivity of peripheral regions to core domestic markets and develop trade agreements could help mitigate the adverse effects of trade barriers and reduce regional disparities within countries.

Keywords: Market Potential, Economic Geography, Regional Development, Core-Periphery
JEL Classification: F15, O18, R11

1.1 Introduction

The core-periphery divide has significant consequences, giving rise to welfare concerns related to unequal access to resources such as education and healthcare, as well as limited employment opportunities. These disparities can extend to the rise of protest movements and voting decisions cleavage, ultimately shaping a nation's political landscape. Among recent prominent examples are Brexit in the United Kingdom, the election of Trump in the United States, and the rise of far-right votes in various European countries. They have shed light on the deep-seated frustrations arising from regional disparities and have highlighted the sense of marginalization and resentment felt by individuals who perceive that urban elites are favored over citizens in rural and peripheral regions (Abreu and Öner, 2020; Dijkstra et al., 2020; Rodríguez-Pose et al., 2020).

A striking fact arises from the evolution in the dispersion of regional GDP per capita within countries. On a sample of more than 100 countries, over 60% of the countries have witnessed a surge in regional disparities from 1995 to 2010, with Gini index growth scaling up to 30%.[1] This divergence extends its footprint across Europe, the Americas, and Asia, underscoring a compelling trend toward growing developmental gaps among regions within nations. The present paper delves into explaining these regional disparities, with a particular focus on the core and periphery divide.

Understanding the causes of regional development disparities is crucial to reduce inequalities within countries and foster social cohesion. Among these factors, proximity to markets has been suggested as a key factor of development. Proximity to demand enables firms to reduce transportation costs when shipping their goods. As a result, they can offer higher wages to their workers, leading to the development of the firm's location as the population becomes wealthier. This concept is captured in the *international trade wage equation* developed by Fujita et al. (1999).

(How much) Does proximity to markets matter in explaining intranational regional development disparities conditional on regional productivity? This paper contributes to the ongoing debate in the New Economic Geography literature by addressing broader questions. It explores the consistency of this effect across countries, the robustness of different measures of market access, and the differential impact on core and peripheral regions. Importantly, the study accounts for productivity-related factors that may also have an influence on regional disparities, including education and population density among others. The study also investigates the distinct effects of proximity to domestic and foreign markets on core and peripheral regions.

Despite the growing concern for peripheral areas being left behind, literature has shown limited exploration of the distinct effects of market access on the development disparities between core and peripheral regions (Brülhart et al., 2004; Brülhart, 2006; Brülhart et al., 2020). In particular, Baum-Snow et al. (2020) are the first to clearly estimate the heterogeneous market access elasticity coefficients on GDP, categorizing prefecture regions between primate and non-primate.

A noteworthy subset of research in this topic is dedicated to the concept of the *border*

[1] See Figure 1.2 to illustrate the point. The computed Gini index represents the dispersion of income per capita among regions within each country. The Gini index ranges from 0 to 1, where 0 represents perfect equality, meaning every region has the same income per capita, and 1 represents perfect inequality, where one region has all the income per capita of the country.

shadow, which characterizes the lower levels of development experienced by regions located in close proximity to international borders (Redding and Sturm, 2008; Adam et al., 2023; Bonadio et al., 2023). These regions often contend with trade barriers, limited market access, and reduced economic opportunities due to their location near borders. This paper aims to contribute to the discussion by conducting an analysis of the core-periphery divide and the differential impact of market access.

These questions are crucial in the context of increasing integration and interdependence among countries and regions, where foreign markets and internal transportation infrastructures play a substantial role in shaping subnational development. This integration and interdependence have been pivotal in recent events, including the Brexit referendum and voting decisions in the United States and various European countries. By examining the relationship between market proximity and regional development, policymakers can develop more targeted policies to foster sustainable and inclusive growth.

Empirical tests of the international trade wage equation of Fujita et al. (1999) emerged in the last two decades, originally focusing on inequality across countries (Redding and Venables, 2004; Fingleton, 2008; Head and Mayer, 2011; Jacks and Novy, 2018). Proximity to wealthy countries is consistently related to their development level. However, considering a country as a market is a substantial simplification. The vast literature of firm location theory together with the core-periphery model developed by Krugman (1991) justify this argument. The concentration of economic activity is often observed in specific locations, commonly referred to as the "core." Typically, core regions are home to a megapolis, and experience vastly different conditions compared to other regions within the same country (Baldwin and Martin, 2004; Redding, 2022).

In the last decade, the focus of the falsification test of the international trade wage equation has switched from cross-countries inequality to within-country inequality. The literature on subnational development emphasizes the diversity of regions within a country and recognizes that subnational approaches can be more effective than national approaches for promoting development. Regions can also be a more relevant unit of analysis for representing markets and understanding the impact of market proximity. For these reasons, the present paper adopt a regional-level position. Due to data availability of development measures and covariates at the regional level, the literature has focused on single-country analysis. Papers have studied the case of the United States, European countries, and more recently, emerging and developing countries. The wider geographical scope of analysis in the literature is found to be at the European level using the rich data produced by European institutions (Niebuhr, 2006; Head and Mayer, 2004; Brülhart et al., 2004; Breinlich, 2006; Head and Mayer, 2006; Brakman et al., 2009; Bruna et al., 2016).

The generalization of the statement to the entire world is hindered by the scarcity of data on regional development measures and covariates, as well as the lack of data required for computing sophisticated market access indexes. The scarcity of data on physical infrastructure for each country and the underutilization of the gravity trade literature, which examines the impact of different distance measures on trade, contribute to these limitations despite their significance as fundamental components for market access measures. This paper aims to address these gaps by making four significant contributions to the literature.

The first contribution of the paper is to provide a falsification test of the international trade wage equation at the regional level within countries and generalize the test to the whole world using a complete set of countries. For robustness, I control for regional development

covariates which proxy for regional productivity, such as population density, education, natural resources, and climatic conditions. The study relies on the large regional dataset built by Gennaioli et al. (2013), which gathers information on Gross Domestic Product (GDP) and covariates for more than 1,500 subnational regions in 107 countries. The dataset covers about 70% of the world's surface and more than 90% of its GDP for the year 2005. For robustness checks, I use a regional panel dataset of about 1,000 regions in 72 countries from Gennaioli et al. (2014) with similar information.

The second contribution is the computation of the index of proximity to markets, the *market potential*,[2] which summarizes the market capabilities of all regions, weighted by their geographical proximity. An effort has been made to fully consider international interactions from trade building a gravity-based index. The proximity function takes into account factors such as common language, national contiguity, colonial ties, common currency, and regional trade agreement as determinants that facilitate closer trade relations between countries and regions, even when they are physically distant. Additionally, since "around 80 per cent of global trade by volume and over 70 per cent of global trade by value are carried by sea and are handled by ports worldwide" as reported by (The Review of Maritime Transport 2018, UNCTAD), I consider international distances between regions through land and maritime shortest paths passing by world ports.[3] Considering the predominant role of transportation infrastructures and geographic limitations allows to recover the travel path of commodities as closely as possible.

The third contribution is on the investigation of the differences in the elasticity coefficients to market potential between core and peripheral regions and how differently they are affected by proximity to foreign markets.[4] To this matter, a classification of regions is done using k-means clustering conditional on the regional economic production. Each country is split into 3 groups of subnational regions: the core, the semi-periphery and the periphery. While core regions benefit from a better access to larger markets, peripheral regions face higher transportation costs,[5] making it more difficult for them to attract investment and develop competitive industries. By employing this classification of regions, the analysis is better equipped to consider the regional variation within countries and gain a better understanding of the specific challenges faced by peripheral regions, specifically according to proximity to domestic and foreign core regions.

The fourth and final contribution lies in a comprehensive examination of the mitigation effects of proximity to markets, specifically focusing on centrality to core markets, distinguishing between domestic and foreign contexts. Integrating centrality measures into the regression analysis not only amplifies the variability in the explanatory variable but also allows for a meaningful contribution to the border shadow literature (Redding and Sturm, 2008; Adam et al., 2023; Bonadio et al., 2023). Additionally, differentiating between centrality to core foreign markets with and without free trade agreements underscores their role in reducing trade barriers, which can be particularly important for regions in the periphery of the economic ac-

[2]The concepts of market potential and market access are used interchangeably throughout the paper.

[3]The title *Navigating the Geography of Regional Disparities* is justified by this aspect of the paper. The term "navigating" is used to refer to the maritime routes taken by ships between world ports, which serves as a metaphor for the investigation of the determinants of subnational disparities.

[4]The initial endeavor in this area was conducted by Baum-Snow et al. (2020). They undertook a classification of prefecture regions based on population size, subsequently identifying primate prefectures. They analyzed the differences in market access elasticity coefficients on GDP between primate and non-primate prefectures.

[5]See Table 1.16 for evidence of lower market access for the periphery.

tivity. The effect of improved centrality to foreign cores with a trade agreement is investigated using the panel subsample.

The analysis is first performed in cross-section to gather as many observations worldwide as possible. The country fixed-effects allow to control for national unobservable variables, as well as to compare the regional observations with their national average for each country. I address endogeneity issues with instrumental variables using a centrality index and a spatial lagged market potential as instruments. The former summarizes the central position to others, and the latter the proximity to foreign markets. Finally, to further test the robustness of the results, the analysis is conducted in panel on a smaller sample of countries.

Results show that proximity to markets is a strong and robust determinant of subnational development all around the world. A 1% increase in the regional market potential is associated with an increase in GDP per capita about 0.1% with respect to countries' average. This elasticity coefficient is similar to those found in single-country analysis (Brakman et al., 2004; Pires, 2006; Mion and Naticchioni, 2009; Fally et al., 2010; Hering and Poncet, 2010; Kosfeld and Eckey, 2010; Paredes, 2013), as well as in European regional analysis (Niebuhr, 2006; Head and Mayer, 2006; Brakman et al., 2009). Conversely, a lower coefficient was identified in comparison to those obtained through country-level analysis (Head and Mayer, 2011), which can attributed to the relatively lower variation in development levels within countries when compared to the variation across countries, or to a more effective control for omitted variables bias. Additionally, the access to ports is revealing to be a powerful determinant of both higher market potential and higher income per capita.

Upon further investigation, heterogeneous elasticity coefficients of regional development to market potential are observed, revealing distinct impacts based on regions' GDP levels. Core regions benefit more from market proximity compared to the (semi-)periphery, with peripheral regions exhibiting a coefficient that is 0.02 percentage points lower than that of core regions. While the proximity to foreign regions has no significant effect on core regions, it can adversely affect regions in the (semi-)periphery. Notably, among peripheral regions, those with higher centrality to foreign cores experience lower levels of development within the same country.

This result is consistent with the findings of Redding and Sturm (2008), who showcase a decrease in population at the Iron Curtain border following the division of Germany after World War II. The establishment of the new border significantly diminished market access for the cities located along this border. With a global pool of countries, Adam et al. (2023) and Bonadio et al. (2023) demonstrate that subnational regions located at international borders, tend to have lower income per capita levels due to their low centrality to domestic markets and international trade barriers.

Results show that proximity to domestic cores exhibits positive effects on regional development while proximity to foreign cores without a trade agreement has a negative impact on peripheral regions. A discussion on firms selection attempting to explain the latter result is provided. However, the effect of proximity to foreign cores with a free trade agreement is interestingly not significantly different from zero. This suggests that the existence of a FTA can help alleviate the negative impact of close proximity to foreign core markets, as supported by the findings of Adam et al. (2023) and Bonadio et al. (2023). However, the panel data analysis reveals that when countries establish a trade agreement and enhance their connections to foreign cores through FTAs, there is no significant expected increase in a region's GDP per capita.

Results highlight the concern for the development divide between core and periphery regions, where the latter tend to have lower income per capita, market potential, and growth rate. To reduce regional disparities and bridge the development gap between core and periphery regions, one possible approach following the present paper's results is to enhance subnational integration by improving national connectivity through transportation infrastructure. Ensuring better access to core national regions may prevent the periphery from being left behind and being excessively impacted by a too close proximity to foreign cores. However, this approach may not be suitable for low-income countries, as indicated by the findings across various country samples. It may pose a risk of regional economic activity and population shifting away from the periphery towards core regions. Conversely, policies geared towards enhancing trade relations and forging free trade agreements with foreign nations can yield advantages for peripheral regions. Simultaneoulsy, policymakers should also work on boosting economic activity in these peripheral regions to prevent economic activity from moving to the core areas.

The paper is organized as follows. Section 1.2 provides a review of the existing literature on regional development and market potential and highlights the gaps in the literature that this paper aims to fill. Section 1.3 introduces the regional market potential indexes used for the analysis. Section 1.4 presents the data. Section 1.5 presents the descriptive statistics. Section 1.6 develops the empirical method. Section 1.7 shows and interprets the results. Finally, Section 1.8 concludes.

1.2 Literature

1.2.1 Wage Equation and Regional Development

Since the emergence of the New Economic Geography literature following the eminent works of Krugman (1980) and Krugman (1991), accessibility to markets has been showed to have a strong impact on the spatial distribution of economic activity. Fujita et al. (1999) have developed a full general equilibrium model with international trade and monopolistic competition, composed by agricultural and manufacturing sectors, which relies on labor immobility. Firms operate under increasing returns to scale and produce differentiated products. From profit maximization, the model express the increasing relationship between factor incomes and market access. This relationship is referred as the *international trade wage equation*.

In particular, a special focus is given on the model developed by Redding and Venables (2004). Redding and Venables' model, which extends the Fujita model by including transport frictions in trade and intermediate goods in production, shows that geographic location affects per capita income via flows of commodities, production factors, and ideas. Distant countries incur trade barriers such as transport costs and cultural differences, leading to reduced market access for exports and imports and lower maximum wage firms can afford to pay due to their zero-profit condition.

Head and Mayer (2011) developed a similar intuition based on the gravity equation for bilateral trade flows and market clearing conditions. The focus has been made on differences in wages, as labor is assumed to be the immobile factor in these theoretical models. However, the wage equation can be generalized to all immobile inputs. Both model imply that distance to markets affects the equilibrium factor prices and so, rises cross-countries income inequality.

The empirical regional-level analysis has widely been developed in the last decade. The

literature has focused on single-country analysis due to data availability, such as the United States, European countries, and more recently, emerging and developing countries.[6] The wider geographical scope of analysis in the literature is found to be at the European level (Niebuhr, 2006; Head and Mayer, 2004; Breinlich, 2006; Head and Mayer, 2006; Brakman et al., 2009; Bruna et al., 2016). The statement has not yet been generalized to the whole world, which is the intention of the present paper.

The literature on subnational development distinguishes between two types of geography that characterize regions and cities: first-nature and second-nature. The former refers to regional natural endowment, such as proximity to water, altitude or climate. The latter refers to the interactions a region can entertain with other regions, depending on its location with respect to them. The literature has argued that the attractiveness of a region is best described by intra-regional characteristics (Combes et al., 2005; Brakman et al., 2009).[7] The present paper is going to question this statement.

In particular, my analysis was inspired by Gennaioli et al. (2013)'s, who investigate the determinants of regional development according to human capital and institutional quality, controlling for first-nature geography. They attest the primary importance of human capital for both regional and national development, while the institutional quality is only significant at the country level. While an effort has been made regarding the first-nature geography dimension, they have omitted the spatial interdependence of regions. Nevertheless, they considered proximity to the coast, which can be used as a proxy for proximity to foreign markets. The variable is showed to have a highly positive significant impact on the regional development levels.

The present paper reevaluates the determinants of regional development with a proper measure of accessibility to markets, referred as market potential, controlling for first-nature geography and education.

1.2.2 Market Potential

Market potential measures the proximity to demand from a given location and represents the attractiveness of a region based on its geographic position and that of its potential trading partners. It reflects a region's potential to trade with others and access to different markets. It is calculated by adding the economic size of all other regions, weighted by the distance that separates them. Regions with high market potential are those that are close to wealthy regions with high demand and market capacities, which allows for low transport costs and higher trade volumes. In contrast, periphery regions, far from the demand, experience low market potential.

[6] Single-country studies of the effect of market access on regional development have been conducted for the United States (Hanson, 2005; Fallah et al., 2009; Donaldson and Hornbeck, 2016; Hornbeck and Rotemberg, 2021), Italy (Mion, 2004; Mion and Naticchioni, 2009; A'Hearn and Venables, 2013; Daniele et al., 2018), Spain (Pires, 2006; López-Rodríguez et al., 2008), Germany (Brakman et al., 2004; Kosfeld and Eckey, 2010) or Finland (Ottaviano and Pinelli, 2006). Other studies have focused on emerging and developing countries, such as Mexico (Chiquiar, 2008), Chile (Paredes, 2013), Brazil (Fally et al., 2010; Hering and Paillacar, 2016), Turkey (Karahasan et al., 2016; Özgüzel, 2022), China (Hering and Poncet, 2010; Baum-Snow et al., 2020; Zou et al., 2021), India (Donaldson, 2018; Bonadio, 2022).

[7] Studying the case of Chile, Paredes (2013) claims that spatial wage inequality may be better explained by amenities than by market access in low income countries. However, some papers demonstrate that the wage equation can hold also in developing and emerging economies (Chiquiar, 2008; Fally et al., 2010; Hering and Poncet, 2010; Karahasan et al., 2016; Özgüzel, 2022).

The internal market of a region also determines its level of market potential. The local market potential depends on its own market capacity and its internal distance as a proxy for internal transport costs. The wealthier the market, the higher the region's own demand, and the higher its market potential. But the larger its land area, the higher the transport costs faced by firms to sell their commodities on average, and the lower the local market potential.

Three different computation methods of the index are developed in the literature: (1) the simple index, as expressed in Harris (1954), (2) the gravity index and (3) the infrastructure index. The first method is a simple function summing the income of the geographical entities, weighted by the inverse of the distance between them (Mion, 2004; Hanson, 2005; Mion and Naticchioni, 2009; López-Rodríguez and Andrés Faíña, 2006; Fallah et al., 2009). Despite the relevant effort of computing market potential indexes which may better translate market accessibility forces, the predictive power of this index has been shown to be as high as other more sophisticated indexes (Daniele et al., 2018).

The second computation method relies on the estimation of the international trade gravity equation and uses other variables of proximity to weight the markets' income, as well as price indexes obtained from exporter and importer fixed effects. This method was adopted by Redding and Venables (2004) and Head and Mayer (2011) for computing indexes at the national level. At the regional level, price indexes has hardly been computed from gravity-based fixed effects since intra-national trade data between sub-national entities is not commonly observable. Some authors have done so in the case of Brazil (Fally et al., 2010; Hering and Paillacar, 2016) and China (Hering and Poncet, 2010). In addition to physical distance between regions or cities, they include other type of proximity measures such as contiguity, common language and regional trade agreement.

The third method for market potential index computation usually relies on the expression of the simple market potential index developed by Harris (1954), using distance measures considering physical infrastructures. A first attempt was made by Hanson (2005), considering a "hub-and-spoke" geodesic distance function, which assumes that goods transported from one county to another must pass through a transportation hub located in the home state of the originating county. Since, road and rail distance has been taken into account (Pires, 2006; Ottaviano and Pinelli, 2006; Paredes, 2013; A'Hearn and Venables, 2013; Karahasan et al., 2016), as well as travel time by car (Brakman et al., 2004; Niebuhr, 2006; Baum-Snow et al., 2020) or train (Zou et al., 2021). Other studies have incurred freight transportation costs as well. For example, Donaldson and Hornbeck (2016) and Hornbeck and Rotemberg (2021) compute average price per ton of transported goods by available transportation routes, using rail and waterways distance. They also use different estimates of elasticity to distance rather than the -1 coefficient used in the simple market potential index.

Few papers integrate the role of international markets in their single-country analysis. An even smallest amount considers the predominant role of ports and maritime transport networks (Ducruet, 2020). In the case of China, Baum-Snow et al. (2020) considers road travel time from each prefecture city to the nearest of the nine largest international ports by value of shipments, as well as the average cost of shipment. In the case of India, Bonadio (2022) estimates trade costs related to ports' access via roads as well as ports' quality, in order to evaluate the relative importance of both infrastructures in shaping international market access of regions. The worldwide scope of my analysis does not allow me to integrate ports' quality, neither inland transportation networks in the market potential indexes. However, an effort has been made to give structure to the function of physical distance.

In this paper, the gravity-based approach is used, with an effort towards the infrastructure approach. I compute two indexes with different specifications of the distance function. Both indices integrates cultural, historical and economical proximity, such as done in the literature using the simple and the gravity index, but different measures of physical distance. The first index uses the haversine distance function, while for the second index, the distance function is provided with more structure, considering the inland shortest path between regions and ports, as well as the overseas shortest path between ports. This distance function considers the world's geography by integrating the delimitation of land and sea spaces and the predominant role of ports in international trade. Next subsection presents the regional market potential indexes.

1.3 Regional Market Potential

To compute the potential to trade of each region, I assume a complete network of regions, where they are supposed to be all interlinked. The intensity of their potential interactions is proportional to their market size weighted by their proximity. The market potential index is expressed as follows:

$$\text{MP}_i = \underbrace{\sum_{j \neq i} y_j \frac{\tilde{y}_j}{\tilde{y}_{\text{max},c_j}} \tau_{ij}}_{\text{NLMP}_i} + \underbrace{b \times y_i \frac{\tilde{y}_i}{\tilde{y}_{\text{max},c_i}} \tau_{ii}}_{\text{LMP}_i} \tag{1.1}$$

where y_j is the GDP of region j, $\tilde{y}_j$ is the GDP per capita of region j, $\tilde{y}_{\text{max},c_j} = \max_{1 \leq k \leq n} \tilde{y}_k$ is the maximum regional GDP per capita among observations within country c_j, τ_{ij} represents the trade costs between regions i and j, τ_{ii} is the internal transport costs in region i, and b is the regional border effect.

The market potential index is composed by two components: the local market potential and the non-local market potential. The former is the second term on the formula's right-hand side, composed by the income information of region i and the weight ranking his hierarchy position among the richest regions in term of GDP. The regional border effect b is giving weight to the internal demand of region i[8]. The intra-regional transportation costs are expressed as follows[9]:

[8]The regional border effect is very rarely estimated in the literature due to the lack of trade flows data between regions of a same country. To proxy for this border effect, I compute b as the average ratio between the maximum domestic market potential on the maximum local market potential among regions in each country, with domestic market potential being the regional market potential with respect to all regions in the same country and the local market potential being the very own market potential of a region, excluding other regions. This ratio forces the local market for a region to weight in the total domestic market, since local demand is also important. I find that $b = 1.5$ for $\text{MP}^{(h)}$ index, and $b = 1.8$ for $\text{MP}^{(s)}$ index. As a comparison, Coughlin and Novy (2021) have estimated states' border effect to be about 1.5 on average in the case of the United States of America for balanced sample over the years 1993, 1997, 2002, and 2000. Hoping that future research will attend to estimate the proper regional border effect within countries for a wider set of countries.

[9]Regions are assumed to be circle-shaped, with the capital city at their center. Then, the average distance from the center of the region and its borders is calculated as a proportion of its radius, with the radius computed as the square root of the land area divided by π. The radius is weighted by 2/3, assuming that the economic activity is relatively concentrated toward the center (Head and Mayer, 2011). Hence, this distance function is an approximation of the average distance between every two cities in region i.

$$\tau_{ii} = \left[\frac{2}{3}\sqrt{\frac{\text{area}_i}{\pi}}\right]^{-1} \tag{1.2}$$

The non-local market potential NLMP_i - first term on the right-hand side of equation 1.1 - corresponds to the sum of the GDP of all regions j different to i, which is multiplied by a weight indicating how well is ranked its income per capita compared to the richest region, i.e. the one with the highest level of GDP per capita.[10] Hence, if j is the richest region of the sample, the coefficient will be equal to 1. Otherwise, it belongs to]0;1[. This ratio gives more weight to richer regions and proxies for the price indices. The closest to 1 the ratio, the higher region's j market power. This component gives the market capacity of region j and is reduced by the trade costs, τ_{ij}, that it faces to import commodities from region i.

For the latter variable, I consider two specifications. The first one is expressed as follows:

$$\tau_{ij}^{(1)} = \text{dist}_{ij}^{(\text{haversine})^{\hat{\beta}_1}} \times \left[\mathbb{1}_{c_i \neq c_j} e^{\hat{\beta}_2 \mathbb{1}_{\text{language}_{ij}} + \hat{\beta}_3 \mathbb{1}_{\text{contiguity}_{ij}} + \hat{\beta}_4 \mathbb{1}_{\text{colony}_{ij}} + \hat{\beta}_5 \mathbb{1}_{\text{rta}_{ij}} + \hat{\beta}_6 \mathbb{1}_{\text{currency}_{ij}}} + \mathbb{1}_{c_i = c_j} e^{\hat{\beta}_7}\right] \tag{1.3}$$

with $\text{dist}_{ij}^{(\text{haversine})}$ the physical distance between regions i and j, computed as the haversine distance, determined by the great-circle distance between two geocodes, which proxies for transportation costs. $\mathbb{1}_{\{c_i = c_j\}}$ and $\mathbb{1}_{\{c_i \neq c_j\}}$ dummies indicating whether region j belongs to the same country than region i. $\mathbb{1}_{\{c_i = c_j\}} e^{\hat{\beta}_7}$ corresponds to the national border effect.[11] It reflects the fact that a pair of regions belonging to the same country is more prone to trade together than with a region abroad, due to more similar consumption preferences, easier communication and the absence of tariffs and customs.

The set of dummy variables equal interacted with the dummy $\mathbb{1}_{\{c_i = c_j\}}$ are equal to one if countries c_i and c_j of regions i and j respectively share a common border, colonial ties, a regional trade agreement and the same currency, as well as if regions i and j share a common language predominantly spoken. These latter variables reflect the cultural and economic proximity between two regions from different countries, which alleviates the effect of distance between them. When $\mathbb{1}_{\text{language}_{ij}}$, $\mathbb{1}_{\text{contiguity}_{ij}}$ and $\mathbb{1}_{\text{colony}_{ij}}$ dummies equal one, they represent lower communication costs, more similar preferences and long term trade, historical and political relationships, which are supposed to increase trade volumes between two regions. $\mathbb{1}_{\text{rta}_{ij}}$ dummy indicates whether the two countries have a trade agreement that lowers trade costs, and $\mathbb{1}_{\text{currency}_{ij}}$ represents the ease for two countries trading with the same currency, as they avoid buying foreign exchange.

The distance between two countries is usually computed as the haversine distance. However, it is a basic measure of distance. This paper provides a better measure of geographical distance, which aims at following as closely as possible the travel undergone by goods from a territory to another, distinguishing the world's areas covered by land and ocean[12]. The new

[10] This ratio is included in the index in order to integrate regional competitiveness. To compute the index with a gravity-based approach, importer and exporter fixed effects are usually included. However, the international trade data used in this paper rely on total bilateral trade flows. Thus, trade flows are not directed - it is not possible to identify an importer and an exporter - and are not at the regional level. Thus, including importer and exporter fixed effects is not relevant.

[11] The literature has estimated $\hat{\beta}_7 = 1.96$ on average (Head and Mayer, 2014).

[12] In 2002, sea transport accounted for 44% of goods, measured in value, traded between the EU and the rest of the world. Measured in volume, the share was about 78%. Regarding road and rail transportation, the share

distance measure is built considering port access[13] for trade that might be undertaken overseas, as well as the shape of land masses, so the path between two regions can get around bodies of water.

To do so, I first collect data on world ports geolocation, and use high-resolution geography data from Wessel and Smith (1996), which provides boundaries between land masses and ocean. From the geographical polygon, I give different values for land masses and ocean and rasterize the map in order to compute shortest paths between each pair of points. The shortest path is computed following the grids of the maps between these points. I first compute the shortest path between each pair of regions by land, then between each pair of region-port by land, and finally, the shortest path between pairs of ports by sea (see figure 1.8 in the appendix for a visual example of the computation). The dataset results in distances computed between each pair of regions, with their associated closest port[14], combining distance by land and sea.

Then, the shipment distance function is computed as follows:

$$\text{dist}_{ij}^{(\text{shipment})} = \begin{cases} \kappa_{io}^{\gamma_1}\kappa_{od}^{\gamma_2}\kappa_{dj}^{\gamma_3} & \text{, if } \kappa_{od} \neq 0 \wedge \frac{\kappa_{io}+\kappa_{dj}+\kappa_{od}}{\kappa_{ij}} < 2 \\ \kappa_{ij}^{\gamma_4} & \text{, otherwise} \end{cases} \tag{1.4}$$

where κ_{od} corresponds to a proxy for the ocean shipping costs. It is the shortest path between the origin port o and the destination port d. Variable κ_{io} is the distance from the region center of region i to its closest port o, and κ_{dj} is the distance from the region center of region j to its closest port d. Hence, κ_{io} and κ_{dj} is the distance traveled by land, and κ_{od} is the distance traveled by sea. I denote by κ_{ij} the distance between regions i and j by land, without passing by any port. If regions i and j are not connected by land, I define $\kappa_{ij} = \infty$. Coefficients γ_1, γ_2, γ_3 and γ_4 are expected to be negative.

The maritime distance function structure $\kappa_{io}^{\gamma_1}\kappa_{od}^{\gamma_2}\kappa_{dj}^{\gamma_3}$ is similar to the one of Bonadio (2022), but he rather uses a port quality measure instead of the maritime distance between ports. Port quality is an important factor of trade costs, however I do not have enough information to estimate it. Moreover, considering maritime distance between ports is relevant as transportation costs also increase with distance on the sea.

The inland and overseas shortest path between regions and ports are computed as straight lines and curves, assuming a continuous space with continental borders constraints. I do not consider physical transportation routes and real costs in the measure as it has been done by some papers in the literature presented in the previous subsection. To do so with a worldwide scope is not possible since this data is not available for most of the countries. The method used in this paper still made progress in considering the actual world geography.

Equation 1.4 integrates a condition of convenience with respect to sea freight. It considers if it is impossible to travel from i to j by land due to non-continuous land area, and if the whole distance between i and j by sea is not larger than twice the land distance between the two, if and only if the maritime distance does not equal zero. Indeed, two regions or countries will not find any interest to reach the same port to trade between each other. The ratio of

was about 30% in value and 13% in volume (Source: Statistics Explained - Eurostat, 2021). As air transportation can be approximated by the haversine distance function, and that it represents a lower share of traded goods as measured both in value and volume, it is not particularly considered it in this paper.

[13]The importance of ports on trade performance has been highlighted in the literature on trade facilitation and logistics (Limão and Venables, 2001; Blonigen and Wilson, 2008; Portugal-Perez and Wilson, 2012).

[14]Bonadio (2022) shows however that firms do not always use the closest port, but incur additional internal costs to reach a port of higher quality.

distances lower than 2 assume regions to use land transportation if the distance travelled by sea is too important with respect to the land distance between two regions, even if maritime transportation is known to display scale economies in the whole transportation costs.

I define $\mathbb{1}_{\text{maritime route}}$ as the convenience condition to ship goods overseas. It equals one if the condition is respected, i.e. if it is convenient for exporters to send its good overseas, zero otherwise. Then, the alternative measure of trade costs becomes:

$$\begin{aligned}\tau_{ij}^{(2)} \quad = \quad & \left[\kappa_{io}^{\hat{\gamma}_1}\kappa_{od}^{\hat{\gamma}_2}\kappa_{dj}^{\hat{\gamma}_3}\,\mathbb{1}_{\text{maritime route}} + \kappa_{ij}^{\hat{\gamma}_4}\left(1-\mathbb{1}_{\text{maritime route}}\right)\right] \times \left[\mathbb{1}_{\{c_i=c_j\}}e^{\hat{\beta}_7}\right. \\ & \left. +\mathbb{1}_{\{c_i\neq c_j\}}e^{\hat{\beta}_2\mathbb{1}_{\text{language}_{ij}}+\hat{\beta}_3\mathbb{1}_{\text{contiguity}_{ij}}+\hat{\beta}_4\mathbb{1}_{\text{colony}_{ij}}+\hat{\beta}_5\mathbb{1}_{\text{rta}_{ij}}+\hat{\beta}_6\mathbb{1}_{\text{currency}_{ij}}}\right]\end{aligned} \tag{1.5}$$

with $\tau_{ij}^{(1)}$ the trade costs using the haversine distance as a proxy for transportation costs, $\tau_{ij}^{(2)}$ the trade costs using the shipment distance function which considers land and sea areas, as well as ports location.

As described above, two indexes of regional market potential are computed with two different measures of geographical distance, the first being computed as the haversine distance $\text{dist}_{ij}^{(h)}$, and the other as the overseas shipping distance $\text{dist}_{ij}^{(s)}$. Hence, the analysis will be conducted using the two different indexes of market potential $\text{MP}_i^{(h)}$ and $\text{MP}_i^{(s)}$. Section 1.A presents the methodology for the computation of the indexes. Coefficients of equations 1.3 and 1.5 are estimated to be: $\left[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4\right]$
$= [-1.18, 0.66, 1.12, 1.37, 0.47, 0.79, 1.96, -0.07, -0.96, -0.06, -1.00]$

Next section describes the regional data used for the computation of the indexes and the analysis.

1.4 Data

The regional data used in this paper mainly come from the dataset built by Gennaioli et al. (2013), with more than 1,500 subnational regions, from 107 different countries, out of the 195 recognized around the world, in the year 2005[15]. This substantially large dataset considers a broad range of geographical, educational, cultural and institutional variables to disentangle the determinants of development at the regional level all over the world.

The regional data of Gennaioli et al. (2013) have been collected at the highest administrative level available for each country: regions, provinces, counties and municipalities, and then aggregated the data to regions and provinces. Since Gennaioli et al. (2013) focus on regions, and aim at identifying the determinants of development within countries, they do not include countries with no subnational administrative divisions in the sample.

The dataset contains the Gross Domestic Product (GDP) per capita of regions on the sample, measured in current purchasing-power-parity (PPP) dollars[16]. This variable is used to proxy for regional development as the dependant variable in the analysis. I also use the GDP

[15]The dataset is available on the website of Rafael La Porta.

[16]The dataset built by Gennaioli et al. (2013) gathers regional income data for 107 countries in 2005, drawn from sources including national statistics offices and other government agencies (42 countries), Human Development Reports (36 countries), OECDStats (26 countries), the World Bank Living Standards Measurement Survey (Ghana and Kazakhstan), and IPUMS (Israel). Their measure of regional income per capita is based on value added but they used data on income (6 countries), expenditure (8 countries), wages (3 countries), gross value added (2

as one of the main component of the indexes of market potential. Figure 1.6 in the appendix shows the world map with the regional data available.

The dataset also includes three measures of geography and natural resources which proxy for the first nature geography: temperature, oil production per capita and the inverse distance to the coast. The first two are gathered from the WorldClim database. The temperature in 2005 is proxied by the average temperature during the period 1950-2000.

Then, two proxy variables of the regional human capital are the population size, which comes from the City Population database, and the average years of education of the 15-year-old and older population, collected from UNESCO database. Zero year of school are allocated for the pre-primary level, six additional school years for the primary education and twelve for those who have completed secondary school. For each region, average years of schooling is computed as "the weighted sum of the years of school required to achieve each educational level". The choice of these variables to explain the development of regions and countries is deeply justified in the development and economic geography literature. Thus, they are not discussed in this paper.

The variable of interest, the regional market potential, is based on the GDP of each region, which proxies for the size of their demand size. To get the most realistic market potential indexes as possible, it is best to consider a complete network of every region around the world. However, there are missing data in the data set of Gennaioli et al. (2013) as they did not include countries for which there were no subnational division data. To solve for this problem in the indexes computation, the GDP information of 68 countries is added from the gravity database developed by CEPII. This extrapolation allows to include a hundred more observation of world's markets into the market potential indexes computation.

In order to compute the distance between each region, which proxies for transport costs, the latitude and longitude are needed for each region. To do so, regions are geocoded on R software, by using a geocoding API key with Google Maps Plateform. After having collected latitude and longitude for each region, an expanded dataset that associates every pair of regions is created, and the distance between each of them is computed using the haversine formula for the measure $\text{dist}_{ij}^{(h)}$, and as the shortest path by land and sea for the measure $\text{dist}_{ij}^{(s)}$ presented in section 1.3.

In order to compute the average internal distance of each region, the land area information is used from Statoids.[17] In order to relate for cultural distances between regions, information on the languages spoken in each region is collected from the same source. Data on countries contiguity and colonial ties is collected from CEPII gravity database (Gaulier and Zignago, 2010). Moreover, information on regional trade agreement and currencies at the country level are collected from the country-level dataset by de Sousa (2012)[18]. The dataset covers 199 countries for the time period 1958-2015. The CEPII database also gathers information on trade for all country pairs between 1996 and 2020.[19] I use this data to estimate the trade elasticity coefficients to resistance and facilitator variables from the gravity equation.

countries), and consumption, investment, and government expenditure (1 country) to fill in missing values. They measured regional income in current PPP dollars because they lacked data on regional price indexes. To ensure consistency with the national GDP figures reported by World Development Indicators, they adjusted regional income values so that, when weighted by population, they total the GDP at the country level.

[17] Visit the Statoids website to find the database - Gwillim Law. Administrative Divisions of Countries. McFarland & Company, Jefferson, North Carolina, October 1999.

[18] Visit the website of the author to find the databases.

[19] The transnational trade panel data comes from BACI.

Finally, geodata on global ports from WPI database are collected in order to compute the shipment distance functions.[20] The database contains the location and physical characteristics of 3,630 ports worldwide, as well as the facilities and services that they offer. I select coastal ports of a substantial size,[21] excluding the smallest ports that may not be used for trade. I also exclude river ports,[22] since rivers' recognition make considerably heavier the shortest path algorithm work. Then, I match to each region its closest port. By doing so, 463 ports remain. Figure 1.7 in Appendix **??** shows the matched ports on a map.

To evaluate if the relationship between the market potential and regional development is robust and persistent over time, I gather panel data from Gennaioli et al. (2014) for the same variables present in their cross-section version for the period 1976-2010.[23] Since there is a substantial amount of missing observations in the early years, I restrict the panel to the period from 1995 to 2010, with 5 years interval and adjust some observations of years around years multiples of 5 in case of missing data.[24] The panel data gather observation for 1,064 subnational regions from 72 countries.

1.5 Descriptive statistics

1.5.1 Regional Income Disparities

This section presents the state of regional inequalities within countries that can be observed in the data. Figure 1.1 displays the Gini coefficient which compares 2005's regional GDP per capita within each country in the sample. The coefficient ranges from 0 to 1, with zero-value referring to completely equal distribution of income per capita across regions, and unit-value referring to the extreme situation of one region holding the total national income, and the rest having no income at all. In other words, the greater the Gini coefficient, the greater development disparities in terms of income per capita distribution across regions. On the map, the darker a country, the greater its territorial inequality.

Kenya exhibits the greatest disparity in regional income per capita, with a Gini index of 0.45, followed by 11 other countries, including Indonesia, Thailand, South Africa, Argentina, Panama, Mozambique, Iran, Russia, Peru, China and Vietnam, all located mainly in South America and Asia. These 12 countries have a Gini index of over 0.3. Conversely, Egypt, Azerbaijan, Pakistan, Syria, France, Malawi and Israel show the least disparity between regions, with a Gini index of less than 0.07.[25]

[20] Visit the WPI website to find the database.

[21] WPI has classified ports' size in four groups according to their total area, wharf space and facilities, i.e. large, medium, small and very small. I exclude the very small ports from the dataset.

[22] Frensch et al. (2023) found that in the case of Europe, international rivers have a modest impact on trade.

[23] The original dataset from Gennaioli et al. (2014) contains different levels of definition of administrative regions for 13 countries compared to the first one from Gennaioli et al. (2013). In particular, the second dataset contains smaller regions. Thus, an effort has been made to combine these two datasets. To do so, I identify and gather regions defined at a lower administrative scale in a larger region, as defined in the regional Gennaioli et al. (2013) data. Among them, I sum the GDP, the population size and the production of oil, and I compute the average amount of education years and the average temperature.

[24] For example, if a country has observations for the year 1994-2001-2005-2010, they are reported to the years 1995-2000-2005-2010 so to have homogeneous time periods between observations.

[25] Egypt, Israel, Malawi, and Pakistan, encompassing areas ranging from 20,000 to 1 million square kilometers, face a limitation in the dataset with only 2 to 6 regions defined. This constraint impedes the accurate estimation of a consistent Gini index for these countries.

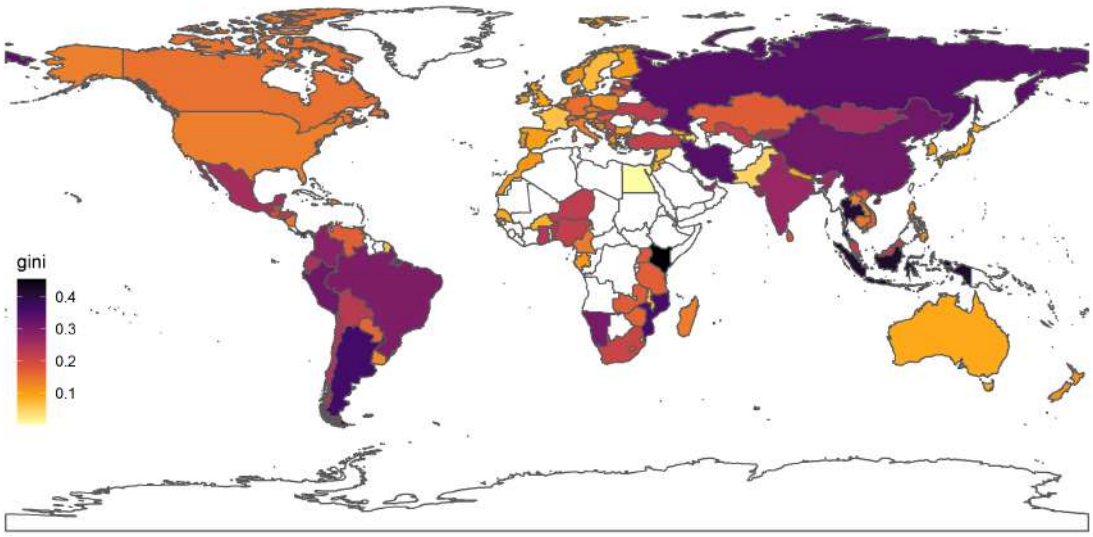


Figure 1.1: Gini index with respect to regional income per capita in 2005

The calculations here display different results than usual analysis of inequality of income, since the focus is given on territorial income per capita inequality rather than household income inequality. Thus, the map in Figure 1.1 is different than the one available on the World Bank website. Their Gini index ranges from 0.25 to 0.65 in 2005, suggesting that household income inequality is, in general, higher than regional income per capita inequality within countries. The most inequal countries according to their calculations were South Africa and Latin American countries, while the most equal countries were in Western Europe. However, we find similar Gini coefficients for Asian countries, around 0.4.

Figure 1.3 in the Appendix depicts the disparity in development between regions within countries, indicated by the ratio of maximum to minimum income per capita. Typically, countries with high Gini indices have a larger ratio of maximum to minimum income per capita.

1.5.2 Regional Market Potential Disparities

As there are great disparities in regional income per capita, there may be so in the regional market potential. Figure 1.4 in the Appendix displays the maximum to lowest market potential ratio within countries, considering the index $MP^{(s)}$, which includes the shortest path distance by land and sea between regions. The income per capita is significantly correlated with the market potential ratios, with a correlation coefficient about 0.7. In other words, higher disparities in regional development should go in hand with higher disparities in trade opportunities.

Russia has the highest ratio of 90, followed by Argentina with a ratio of 32, and Japan with a ratio of 20. These countries have even greater disparities in terms of market potential than in terms of GDP per capita. In Europe, while regional development disparities appear higher in the East than the West, the opposite trend is observed for market potential. Higher ratios of highest-to-lowest market potential indicate significant income per capita dispersion. The wealthiest regions have wealthier neighboring regions, while the poorest regions are located further away. A high ratio indicates strong wealth concentration over space. However, a high ratio can also be due to larger regional area delineation, resulting in greater distances

between regions. This explains the significant disparities between regions in Australia and the United States (where regions are defined as states). African countries display particularly low differences in market potential due to similar foreign market potential indices. Their distance from the world's wealthiest regions is similar.

Table 1.13 in the Appendix presents the descriptive statistics of the variables used in the analysis for the whole 2005's sample, as well as by country income group following the World Bank classification: *High income*, *Upper middle income*, *Lower middle income* and *Low income*.[26] The table indicates that market potential indexes using different distance functions are similar. Additionally, the richer the country, the higher the average regional GDP per capita, education level and market potential indexes.

The table presents interesting differences in the distribution of GDP per capita and market potential across different income groups. In the overall sample, the standard deviations of GDP per capita and market potential are equal, indicating a relatively balanced dispersion of these variables. However, when the analysis is stratified by income level, distinct patterns emerge.

In high and upper-middle-income countries, the standard deviation of GDP per capita is lower than that of market potential. This indicates that, despite substantial heterogeneity in market access contributing to greater diversity in economic development potential, there exists a more consistent distribution of income levels among these nations. This suggests a certain degree of economic convergence or stability in their overall economic performance.

Conversely, low-income countries demonstrate lower standard deviations in market potential compared to GDP per capita. This observation may imply that minimal variations in market access can lead to significant disparities in development levels, potentially highlighting challenges in achieving economic stability or convergence.

1.5.3 Core and Periphery Divide

To investigate regional heterogeneity and concentration of economic activity, I classify regions within each country into three groups based on their Gross Domestic Product: (1) core regions, representing the wealthiest areas, (2) periphery, denoting the least affluent regions, and (3) semi-periphery, encompassing those in between. Utilizing k-means clustering, I assign regions to their respective groups. It is important to note that the number of regions in each group varies across countries. However, every country, with the exception of those with fewer than three regions like Ireland, has at least one region classified in each of the mentioned groups.

[26] **List of high-income countries (# regions):** Australia (8), Austria (9), Belgium (11), Canada (12), Switzerland (25), Czech Republic (8), Germany (16), Denmark (15), Spain (19), Estonia (15), Finland (5), France (22), Great Britain (12), Greece (13), Croatia (20), Hungary (7), Ireland (2), Israel (6), Italy (20), Japan (47), Netherlands (12), Norvege (19), New Zealand (14), Poland (16), Portugal (7), Slovakia (8), Slovenia (12), Sweden (8), United States of America (51). **List of upper-middle income countries (# regions):** Azerbaijan (11), Bulgaria (6), Bosnia and Herzegovina (3), Brazil (27), Chile (13), China (31), Colombia (33), Dominican Republic (9), Ecuador (21), Gabon (4), Iran (30), Jordan (12), Kazakhstan (6), Lebanon (6), Lithuania (10), Latvia (33), Mexico (32), Macedonia (8), Malaysia (14), Panama (9), Peru (24), Romania (8), Russia (80), Thailand (5), Turkey (12), Uruguay (19), Venezuela (23), South Africa (9). **List of lower-middle income countries (# regions):** Armenia (11), Belize (6), Bolivia (9), Cameroon (10), Egypt (2), Georgia (11), Ghana (10), Guatemala (8), Honduras (18), Indonesia (33), India (32), Laos (18), Morocco (14), Moldova (5), Mongolia (22), Nigeria (6), Nicaragua (17), Pakistan (5), Philippines (17), Paraguay (18), Senegal (10), El Salvador (14), Swaziland (4), Syria (13), Ukraine (27), Uzbekistan (5), Vietnam (8). **List of low income countries (# regions):** Benin (13), Burkina Faso (13), Kenya (8), Kyrgyzstan (8), Cambodia (15), Madagascar (6), Mozambique (10), Malawi (3), Niger (8), Nepal (5), Tanzania (21), Uganda (4), Democratic Republic of the Congo (11), Zimbabwe (10).

	(1)	(2)
$\mathbb{1}_{\text{semi-periphery}}$	-0.39^{***}	-0.33^{***}
	(0.04)	(0.05)
$\mathbb{1}_{\text{periphery}}$	-0.62^{***}	-0.43^{***}
	(0.05)	(0.06)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{Upper-middle income country}}$		-0.06
		(0.09)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{Lower-middle income country}}$		-0.01
		(0.09)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{Low income country}}$		-0.32^{**}
		(0.14)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{Upper-middle income country}}$		-0.32^{***}
		(0.11)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{Lower-middle income country}}$		-0.17
		(0.12)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{Low income country}}$		-0.45^{**}
		(0.18)
Num. obs.	1516	1516
Num. groups: code	104	104
Adj. R^2 (full model)	0.93	0.93
Adj. R^2 (proj model)	0.23	0.25

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. The dependant variable is the regional income per capita (in log). Robust standard errors adjusted for clustering on each country are in parentheses. Regressions include country fixed effects.

Table 1.1: Regional Development, the Core and Periphery Divide

Statistics on these clusters are presented in table 1.C. The core cluster counts between 1 and 6 regions, i.e. between 2% and 38% of the total amount of regions within countries. The semi-periphery counts between 1 and 12 regions, i.e. 3% and 67% in the total amount of regions within countries. Finally, the periphery counts between 1 and 66 regions in each country, which represents from 17% to 93% of the whole national territory. Therefore, the world counts much more peripheral regions than core regions. On average, the GDP per capita is higher in core regions compared to semi-peripheral and peripheral regions. This observation also holds true for population density and years of education.

To observe regional disparities of income between core and peripheral regions within countries, I run regressions of income per capita on the different clusters' dummies, including country fixed effects. Table 1.1 show the results. The average income per capita in semi-peripheral and peripheral regions are found to be respectively 39% and 62% lower than that of core regions. When investigating differences between core and peripheral regions in the different countries' income groups, results show that the core-periphery divide is greater for upper-middle and low income countries than high-income and low-middle income countries.

Tables 1.15 and 1.16 in the appendix show the results using education, density and market potential indexes as dependant variables. (Semi-)peripheral regions are shown to be less educated, less inhabited and further from markets than core regions. Additionally, results show that the lower the income group of countries, the higher the disparities in education levels

between core and peripheral regions.

Now that we have showed evidence of the existence of strong disparities within countries, it is also possible to study their evolution using panel data. Figure 1.2 shows the average growth in the Gini coefficient within countries from 1995 to 2010. We observe that 38% of the countries saw the dispersion of income per capita to decrease, with a Gini coefficient average growth ranging from 0 to -10% (from yellow to green areas on the map), while the others saw their disparities between regions to increase, with a Gini coefficient average growth ranging from 0 to 30% (from yellow to red areas on the map). The observed increase in regional disparities suggests a growing tendency towards divergence in the levels of development between different regions.

	Model 1	Model 2
year	0.0254^{***}	0.0179^{***}
	(0.0029)	(0.0035)
year $\times \mathbb{1}_{\text{semi-periphery}}$	-0.0002^{***}	-0.0001^{***}
	(0.0000)	(0.0000)
year $\times \mathbb{1}_{\text{periphery}}$	-0.0002^{***}	-0.0002^{***}
	(0.0000)	(0.0000)
year $\times \mathbb{1}_{\text{upper-middle income country}}$		0.0090^{*}
		(0.0047)
year $\times \mathbb{1}_{\text{lower-middle income country}}$		0.0099
		(0.0094)
year $\times \mathbb{1}_{\text{low income country}}$		0.0288^{***}
		(0.0062)
year $\times \mathbb{1}_{\text{upper-middle income country}} \times \mathbb{1}_{\text{semi-periphery}}$		-0.0000
		(0.0001)
year $\times \mathbb{1}_{\text{upper-middle income country}} \times \mathbb{1}_{\text{periphery}}$		-0.0001^{**}
		(0.0001)
year $\times \mathbb{1}_{\text{lower-middle income country}} \times \mathbb{1}_{\text{semi-periphery}}$		-0.0001
		(0.0001)
year $\times \mathbb{1}_{\text{lower-middle income country}} \times \mathbb{1}_{\text{periphery}}$		-0.0001
		(0.0001)
year $\times \mathbb{1}_{\text{low income country}} \times \mathbb{1}_{\text{semi-periphery}}$		-0.0003^{***}
		(0.0001)
year $\times \mathbb{1}_{\text{low income country}} \times \mathbb{1}_{\text{periphery}}$		-0.0003^{***}
		(0.0001)
Num. obs.	3566	2946
Num. groups: code	58	49
Adj. R^2 (full model)	0.8985	0.8837
Adj. R^2 (proj model)	0.2204	0.2350

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. The dependant variable is the GDP per capita (in log). Robust standard errors adjusted for clustering on each country are in parentheses. Estimations include country fixed-effects.

Table 1.2: Regional Development, Core and Periphery Divide

To investigate differences in development trends experienced by core and peripheral re-

gions, I regress the regional income per capita on the year and interacts it with dummies for semi-peripheral and peripheral regions with country fixed effects. Results are displayed in table 1.2. Conditional on countries' unobservables characteristics, findings reveal that core regions witness annual average growth rate in income per capita about $(\exp^{0.0254} - 1) \times 100 = 2.57$ % on average, while the growth rate experienced by (semi-)peripheral regions is 0.02 percentage points lower. Adding interactions terms with dummies of countries' income group, it is found that the lower the countries' income, the higher the growth rate of their core regions, but also the lower the growth rate of their peripheral regions. These findings suggest a global convergence among core regions but also reveal a widening development gap between core and periphery regions within countries. The higher growth in core regions compared to peripheral regions exacerbates regional disparities over time.

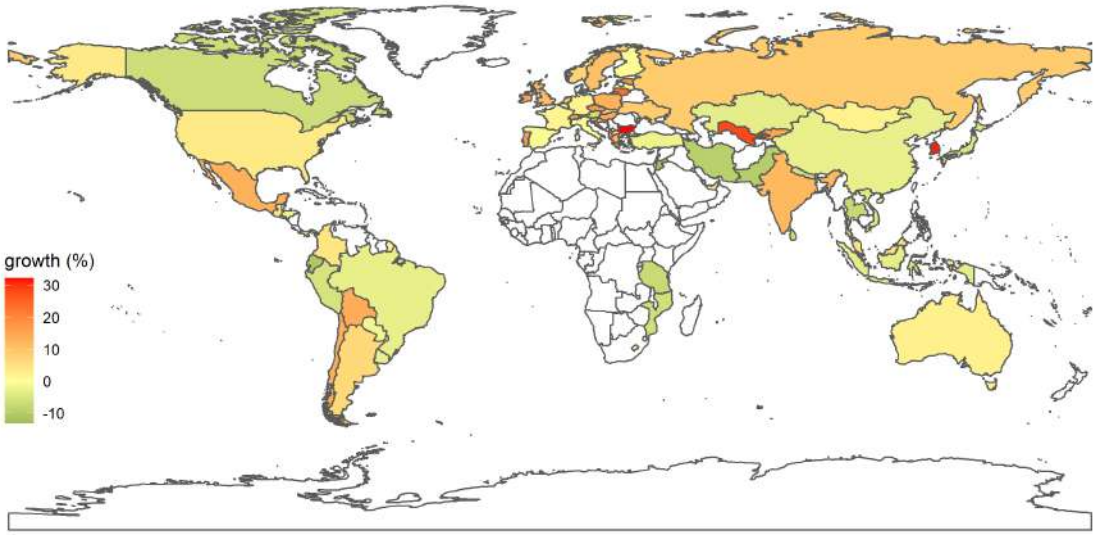


Figure 1.2: Gini coefficient average growth from 1995 to 2010

This increase in regional disparities between core and peripheral regions raises serious equity concerns. While regional disparities can be natural to a certain extent, because of different natural conditions and resources' availability, the rise to persistent and wider disparities highlight a concentration process of economic activity, leaving peripheral areas behind. Indeed, markets favor some places over others. Favored places and firms located there may gain in terms of productivity from the benefits of agglomeration economies, and further grow over time. On the contrary, remote and low productive areas become further remote and less productive, under-utilizing their potential and experiencing retrained growth. This paper aims at testing this assumption, using the market potential indexes and other regional development covariates.

Next section presents the empirical model for the study of subnational development determinants used as a falsification test for the international trade wage equation at the regional level within countries.

1.6 Empirical Model

The effect of market potential on wages can be estimated by transforming the wage equation in the logarithm form. The coefficient estimated will be interpreted as an elasticity. In other words, I am looking for the percentage change in wages according to a 1% change in market potential. Following Redding and Venables (2004) and Head and Mayer (2011), who empirically tested the international trade wage equation at the country level, income per capita is used as a proxy for wages.[27] Following this strategy, I can enlarge the interpretation of the usual wage equation and read the estimated coefficient of market potential as the elasticity of regional development with respect to market potential.

1.6.1 Baseline estimation

To conduct the analysis, the regional income per capita is regressed on market potential and covariates that are known to have a significant impact on regional development within countries. The covariates are geographical variables and human capital proxies. I expect the errors terms of GDP per capita regressions to be correlated with some specific group effect. Each region in a single group may be correlated but independent across groups, i.e. $Cov(u_{c_i,i}, u_{c_i,j}) = \zeta_{ij}^{c_i}$ and $Cov(u_{c_i,i}, u_{c_j,j}) = 0$ for any region $i \neq j$ belonging to countries c_i and c_j.[28] Country fixed effects are included in order to control for the long-run and historical factors that have shaped the economic development of countries and their institutions, aiming at reducing the omitted variable bias.

Unlike Gennaioli et al. (2013), I do not use their institutional quality index at the regional level because it is not found to affect GDP per capita and is limited to too few regions. I assume that institutional quality is uniform within a country and accounted for by using country fixed effects. Using country fixed effects eliminates the between-country variation by demeaning the data within countries. Hence, the results will be interpreted as average effects with respect to countries' average. The following regression is estimated:

$$\ln \text{GDPpc}_i = \alpha_1 \ln \text{MP}_i^{(h,s)} + \sum_{k=1}^{5} \delta_k X_i^{(k)} + \zeta_c + u_i \tag{1.6}$$

where α_1 is the coefficient of interest, $\text{MP}_i^{(h,s)}$ the market potential index computed either with the haversine distance $\text{dist}_{ij}^{(h)}$, either with the shipment distance function $\text{dist}_{ij}^{(s)}$ defined

[27] Assessing development by income per capita is largely adopted in the literature and allows to overcome the lack of available data regarding wages at the regional level worldwide, even if it is generally reasonable to criticise this strategy. Indeed, GDP per capita is not the only measure of development. There exist a large variety of indexes taking in consideration level of education, institutions, health care, etc. However, GDP per capita has a general impact on all of these variables. Indeed, richer countries tend to display higher educational and health care attainments, which essentially participate to the welfare of societies. Thus, GDP per capita may be reasonable to consider as a proxy for development. Further research can be conducted using other proxies for development. A difficulty in doing so may be in the absence of a database with as many regional observations worldwide. For the same reason, Gross Value Added (GVA) per capita is not used as a proxy for wages. This is not a serious limitation as studies that use GVA per capita (Breinlich, 2006; Bruna et al., 2016) find similar results than those using GDP per capita (Head and Mayer, 2006) in the case of Europe. Moreover, in some countries, wages are set at the national level for many production sectors.

[28] The average number of observations per country is about 14.2, with the minimum about only 2 regions (Egypt and Ireland), and the maximum about 80 regions (Russia).

in equation 1.4, ζ_c the country fixed-effect. Component $X_i^{(k)}$ represents the GDP per capita covariates, with k referring to each of the five following variables: temperature, inverse distance to port, oil production per capita, average years of education, and population density.

1.6.2 Robustness

Proxies of market potential

The market potential variables are suspected to be endogenous since their local component is expressed as the local GDP_i. It induces a problem of reverse causality between the regressor and the regressand since income and market potential are simultaneously determined. It possibly creates biased OLS coefficients for market potential. To first solve for this problem, I replace the market potential by the non-local market potential, i.e. $NLMP_i$.[29] Doing so deletes the endogenous local part of the market potential.

However, proxying the market potential MP_i by the non-local market potential $NLMP_i$ may not be sufficient to erase the problem of endogeneity, and the coefficients of non-local market potential indexes could still be biased. As the first Law of Geography by Tobler (1969) states: "Everything is related to everything else, but near things are more related than distant things". To further delete the near information in market potential indexes, market potential indexes are also proxied by the foreign market potential FMP_i. It deletes information on domestic regions. The correlation between market potential and foreign market potential indexes are about 0.55 and 0.64 respectively for the indexes including distance measures expressed as $dist_{ij}^{(h)}$ and $dist_{ij}^{(s)}$.

The non-local market potential $NLMP^{(h)}$ represents about 96% (98% for the $NLMP^{(s)}$ index) of the market potential on average, while the foreign market potential, about 65% (69% resp.). Thus, the omission of the local and domestic part of the market potential is substantial. Regressing the GDP per capita on market potential and covariates, and each of the variables $NLMP_i$ and FMP_i allow us to test whether proxies are good if they are redundant, i.e. unsignificant as the proxied variable is present in the regression. The foreign market potential is found to be a better proxy for the market potential index, as its elasticity coefficients are unsignificant.

2SLS estimations

We saw that the local market potential represents a substantial part of the market potential. Erasing information on the local market may not be the best solution to solve for endogeneity as it induces a problem of omitted variable. There may also be the problem of spatial endogeneity. In order to better solve the endogeneity problem and to keep the local and domestic information of the market potential at the same time, Two-Stage-Least-Squares (2SLS) estimations are conducted.

Redding and Venables (2004) chose the log distance to the three biggest international markets as instruments for their market access index, which are Japan, Western Europe, with Belgium at its center, and the USA. Head and Mayer (2011) chose to use a centrality index, which is computed as the sum of the distances between every countrie. The higher their centrality measure, the closer a country to all the other markets regardless of their market size, and the lower its transportation costs.

[29] Non-local market potential is also often referred as external market potential (Harris, 1954).

Using this latter IV strategy to regions, the instruments are found not to be exogenous. The arising problem comes from the fact that the size of regions has a direct impact on proximity to other markets, as well as on GDP per capita. A smaller region in terms of land area tend to be closer to other markets, since its centroid is naturally closer to others than a vaster region. Thus, its centrality index will be higher, since the sum of the inverse of distances is naturally higher. However, administrative regions that have been designed to be smaller in terms of land area are often the regions with the highest level of GDP and GDP per capita. Centrality is a good instrument for market potential considering countries, but it is not when considering sub-national regions, which are defined for administrative and political purposes.

The IV strategy I am using in this paper is halfway between Head and Mayer's and Redding and Venables'. I define centrality$_i$, the harmonic centrality of each region with respect to the richest market of each country in the sample. The index provides information on the proximity to these markets such as the higher the index, the closer the region to the richest national markets. It may help better capturing the regional heterogeneity in proximity to rich regional markets. The instrument is computed as follows:

$$\text{centrality}_i = \sum_{j(r)} \frac{1}{\text{dist}_{ij(r)}^{(h,s)}} \tag{1.7}$$

with r the richest region in terms of GDP of country j, and $\text{dist}_{ij(r)}^{(h,s)}$ the distance between i and r for each country j, computed as the haversine distance and as the shipment distance function defined in equation 1.4. More particularly, $\text{centrality}_i^{(h)}$ is going to depend on $\text{dist}_{ij(r)}^{(h)}$, and $\text{centrality}_i^{(s)}$ on $\text{dist}_{ij(r)}^{(s)}$. The first stage regression is expressed as follows:

$$\ln \hat{\text{MP}}_i^{(h,s)} = \lambda_1 \ln \text{centrality}_i^{(h,s)} + \sum_{k=2}^{6} \lambda_k X_i^{(k)} + \zeta_c + \epsilon_i \tag{1.8}$$

Another instrument used for market potential is the foreign market potential. Indeed, it is found to not have a significant impact on the regional income per capita, and it is redundant when including it in equation 1.6, i.e. its coefficient is not significant - see the results in section 1.7.2. Hence, another version of the first stage regression is:

$$\ln \hat{\text{MP}}_i^{(h,s)} = \theta_1 \ln \text{FMP}_i^{(h,s)} + \sum_{k=2}^{6} \lambda_k X_i^{(k)} + \zeta_c + \epsilon_i \tag{1.9}$$

Panel estimation

To test whether regional development is robustly explained by market potential, I estimate the model with the panel data, and interact the country fixed effects with years, and denote it as ζ_{ct}. Doing so allows to control for heterogeneity of countries at different points in time.[30] In

[30]Note that the only source of temporal variation for the market potential indices is the GDP information. Literature has shown that ocean shipping costs are falling over time for all countries as improved technologies reduce port time and speed sea travel and larger containers transport larger volumes of commodities. In particular, Hummels (2007) reports that from the 80s to 2000s, shipping costs have been declining steadily. However, the shipment distance function used in the computation of market potential indexes do not vary in this analysis. This omission can be translated in a decrease in the elasticity coefficients to market potential over time. We do not find that the coefficient change (see results from the cross-sectional estimations for 1995, 2000, 2005 and 2010 in table 1.21 in the Appendix).

particular, these country-year fixed effects allows to control for possible change in countries' total factor productivity or institutional quality over time. Coefficient α_1 gives the average effect of a regional market potential deviation from country-year average.

1.6.3 The Core and Periphery Divide

The heterogeneous effect of market potential on regional development is also investigated for different groups of regions identified within each country according to their GDP levels using k-means clustering. The resulting groups of regions are classified as core, semi-periphery and periphery regions. They reflect the results of agglomeration effects described by Krugman (1991). I estimate the following specification:

$$\ln \text{GDPpc}_i = \alpha_1 \ln \text{MP}_i^{(h,s)} + \sum_{g=2}^{3} \alpha_g \ln \text{MP}_i^{(h,s)} \times \mathbb{1}(\Gamma_i = \gamma_g) + \sum_{k=1}^{5} \delta_k X_i^{(k)} + \zeta_c + u_i \quad (1.10)$$

with Γ_i the classification of region i, with $\gamma_g = \{$*core, semi-periphery, periphery*$\}$ the different values of Γ_i. α_1 is the average percentage change of a 1% increase in market potential index from the national average for a core region, ceteris paribus. α_2 and α_3 are the effect of the two interaction terms of the market potential variable and dummies, which are equal to 1 if region i is a semi-periphery or a periphery region respectively, zero otherwise. In other words, they represent the difference in percentage points of the elasticity to market potential between (semi-) peripheral and core regions. In order to investigate potential heterogeneity effects accross countries, I split the sample into four groups based on the level of income of each country: high-income, upper-middle-income, lower-middle-income, and low-income countries. It allows to determine the different degrees of sensitivity to market access for core and periphery regions across different countries' income levels.

Centrality to core markets

Additionally, the study delves deeper into the impact of proximity to domestic and foreign core regions on (semi-)peripheral regions. In order to do this, other measures of centrality are used, which are based on the inverse haversine distance between a region and the national or foreign core regions. Note that based on the k-means clustering of regions according to their GDP levels, countries in the sample contain between one to three core regions. The inclusion of centrality measures in the regression analysis helps provide a clearer insight into the influence of proximity to affluent markets by addressing certain limitations associated with market potential indexes. These indexes tend to aggregate a multitude of values, leading to limited variation between regions within countries. By introducing centrality measures, a more precise understanding of the connection between market access to core domestic and foreign regions and regional development disparities between core and peripheral regions can be attained, thanks to the additional variation they introduce into the regression estimations.

I distinguish centrality to foreign core regions between centrality to foreign core with a free trade agreement and without. Being in close proximity to foreign cores without a trade agreement can impede trade due to national barriers such as tariffs, quotas, and regulations. These barriers restrict the movement of goods, services, and investments, increasing costs and hindering trade for peripheral regions. Despite their geographical closeness, the absence of

a trade agreement limits the potential economic benefits that could arise from proximity to foreign cores, resulting in missed trade opportunities and reduced growth prospects.

I will contrast the findings with those of Adam et al. (2023) and Bonadio et al. (2023) who find that that subnational regions near international borders often exhibit lower per capita income levels and that trade agreements at these borders can mitigate this negative impact by facilitating trade and reducing barriers. Although the present papers' specification regarding proximity to foreign cores without a trade agreement may not directly align with Adam et al. (2023) and Bonadio et al. (2023)'s results, they provide additional insights into the compensatory role of trade agreements in mitigating the adverse consequences of geographical proximity to foreign cores.

The four centrality measures are:

$$\text{centrality}_i^{\text{domestic cores}} = \sum_j \frac{1}{\text{dist}_{ij}} \times \mathbb{1}(\Gamma_j = \text{core}) \times \mathbb{1}(c_i = c_j) \tag{1.11}$$

$$\text{centrality}_i^{\text{foreign cores}} = \sum_j \frac{1}{\text{dist}_{ij}} \times \mathbb{1}(\Gamma_j = \text{core}) \times \mathbb{1}(c_i \neq c_j) \tag{1.12}$$

$$\text{centrality}_i^{\text{foreign cores, FTA}} = \sum_j \frac{1}{\text{dist}_{ij}} \times \mathbb{1}(\Gamma_j = \text{core}) \times \mathbb{1}(c_i \neq c_j) \times \mathbb{1}(\text{rta}_{c_i,c_j} = 1) \tag{1.13}$$

$$\text{centrality}_i^{\text{foreign cores, no FTA}} = \sum_j \frac{1}{\text{dist}_{ij}} \times \mathbb{1}(\Gamma_j = \text{core}) \times \mathbb{1}(c_i = c_j) \times \mathbb{1}(\text{rta}_{c_i,c_j} = 0) \tag{1.14}$$

where $\mathbb{1}(\Gamma_j = \text{core})$ correspond to a dummy stating that region j is a core region in country c_j. $\mathbb{1}(c_i = c_j)$ refers to domestic relationship and indicates that countries of regions i and j are the same if the dummy equals 1, zero otherwise. $\mathbb{1}(c_i \neq c_j)$ refers to foreign relationship where countries of regions i and j are different. If they are different, the dummy equals 1, zero otherwise. Among those foreign relationships, a distinction is done between $\mathbb{1}(\text{rta}_{c_i,c_j} = 1)$ and $\mathbb{1}(\text{rta}_{c_i,c_j} = 0)$, which indicate whether or not the countries of regions i and j have a trade agreement in 2005.

I estimate the following specifications:

$$\ln \text{GDPpc}_i = \sum_{w=1}^{W} \alpha \ln \text{centrality}_i^{(w)} + \sum_{k=1}^{5} \delta_k X_i^{(k)} + \zeta_c + u_i \tag{1.15}$$

$$\ln \text{GDPpc}_i = \sum_{g=1}^{3} \sum_{w=1}^{W} \alpha_{wg} \ln \text{centrality}_i^{(w)} \times \mathbb{1}(\Gamma_i = \gamma_g) + \sum_{k=1}^{5} \delta_k X_i^{(k)} + \zeta_c + \gamma_g + u_i \tag{1.16}$$

with the first one investigating the effect of the centrality measures that contain different weights which indicate whether it relates to centrality to domestic or foreign cores, with and without free trade agreement. The second specification estimate the coefficients for each group of regions, i.e. the core and the (semi-)periphery.

I conduct robustness checks using panel data to further assess the effects of centrality on foreign core regions that have a trade agreement with the region's country of reference. Specifically, I evaluate the impact of signing a trade agreement by computing the centrality index for foreign core regions with a trade agreement for each year of the panel sample (1995, 2000,

2005, 2010) and explore how variations in the index influence regional development. When a country signs a trade agreement with another, the centrality index of its regions increases as I add the sum of the inverse distances to the core regions of the second country. I estimate the following regressions:

$$\ln \text{GDPpc}_{it} = \alpha \ln \text{centrality}_{it}^{\text{foreign cores, FTA}} + \sum_{k=1}^{5} \delta_k X_{it}^{(k)} + \zeta_{ct} + u_{it} \tag{1.17}$$

$$\ln \text{GDPpc}_{it} = \sum_{g=1}^{3} \alpha_g \ln \text{centrality}_{it}^{\text{foreign cores, FTA}} \times \mathbb{1}(\Gamma_i = \gamma_g) + \sum_{k=1}^{5} \delta_k X_{it}^{(k)} + \zeta_{ct} + \zeta_i \times \gamma_g + u_{it} \tag{1.18}$$

where α is the average percentage effect of a 1% increase in the variable $\text{centrality}_{it}^{\text{foreign cores, FTA}}$ due to trade agreement with other countries, and α_g is the specific effect for each group of regions, the core and the (semi-)periphery. ζ_{ct} is a country-year fixed effect and $\zeta_i \times \gamma_g$ a region-cluster group fixed effect. Results are displayed and discussed in next section.

1.7 Results

In this section, evidence is presented on the determinants of regional development, first in cross-section in 2005, then in panel estimations for the years $t = \{1995; 2000; 2005; 2010\}$. A particular focus on the effect of market potential is given. Regressions allow to test the international trade wage equation and to compare coefficients with the structural parameter from the theory.[31]

1.7.1 Baseline estimations

Univariate regressions of GDP per capita on market potential and its proxies are presented in table 1.17 in the appendix. Results show positive and highly significant coefficients for the estimations with the different distance function. In particular, if the average region in a country experience a 1% increase in its market potential, it is expected to increase its GDP per capita by 0.2%. The R-squared statistics indicate that variations in regional $\ln$ GDPpc are explained by between 4% and 6% of the variations in $\ln$ MP within countries. R-squared statistics are lower when considering the non-local market potential indexes, i.e. 0.02. The foreign market potential does not seem to explain regional development. $\text{MP}^{(h)}$ and $\text{MP}^{(s)}$ display similar coefficients among the different specifications.

It is important to note that the elasticity coefficient should theoretically equate to $1/(\beta\sigma)$, where β represents the income labor share, and σ signifies the elasticity of substitution be-

[31] Coefficients may be equal to $1/\beta\sigma$, with σ the elasticity of substitution between varieties and β the income labor share. Following a methodology derived from a model of international trade in differentiated products and monopolistic competition, the elasticity of substitution between varieties has been estimated in the literature. At the aggregate level, it has been found to be around 6 and 11 - see Head and Ries (2001), Lai and Trefler (2002), Romalis (2007). Depending on the industry field, Ahmad and Riker (2020) find that the elasticity of substitution between varieties can range from 1 to 13, with the median around 2 and 3. On the other hand, the income labor share is generally estimated to be between 0.5 and 0.75 - see Guerriero (2019). Assuming these ranges for σ and β, we can expect the coefficient of interest of the analysis to be in the interval $[0.10; 0.33]$, or even in the interval $[0.44; 2[$ for values in the lower bound of σ.

tween varieties. Regarding the income labor share, Reshef and Santoni (2023) conduct estimations and found that it ranged from 0.3 to 0.7, depending on the country, as of 2007.[32] Furthermore, Fontagné et al. (2022) estimate the elasticity coefficient of substitution between varieties at the product level and found values between 5 and 20. The expected α_0 coefficient that we need to find should fall between 0.07 and 0.7. The coefficients revealed in the present paper fall within the expected range, which is reassuring both for validating the theoretical framework and the worldwide falsification test.

Table 1.3 shows the baseline results in cross-section for the year 2005. The estimations are controlled for other geographical variables and human capital proxies, such as *temperature*, known to affect labor productivity, *inverse distance to the closest port*, which is used as another measure for the impact of international trade activities as it can proxy for foreign markets proximity, *production of oil per capita*, as it is a particular endowment that affect development in specific ways, and finally *average years of education* and *population density*, which both play a role as knowledge resources and production means. The covariates tend to reduce the biases of the univariate regressions.

Conditional on time-invariant characteristics of countries, the elasticity coefficient of regional development to market potential is found to be equal to 0.10, and 0.05 for the non-local market potential. In other words, a 1% higher market potential index is associated to a 0.1% higher GDP per capita. These coefficients correspond to high values of elasticity of substitution between varieties. Notice that the non-local market potential as measured by port accessibility and ocean shipment distance function give a relatively lower confidence interval, but the p-value is still about 0.057. The foreign market potential is not found to have a significant impact on the GDP per capita differences within countries. It is also the case when we exclude *inverse distance to the closest port* from the regression. The lack of significance in the foreign market potential effect might be attributed to the limited variation captured by the index within countries, as indicated by the regression results of the index against (semi-)periphery dummy variables - refer to Table 1.16 in the appendix.

Despite the non-significance of the foreign market potential coefficient, accessibility to foreign markets as proxied by the inverse distance to ports is found to be a highly significant determinant of GDP per capita. The coefficient is robustly estimated to be equal to 0.14. The oil production per capita and the average years of education are also found to explain regional differences in development levels within countries. In particular, an additional year in the average level of education in a region with respect to the national average is found to increase the GDP per capita by 28%. This coefficient is lower than the one obtained by Gennaioli et al. (2013) (see table IV, page 134).

It is worth to notice that the effect of education on regional development keeps being much larger than the micro estimates, which are generally of the 0.06-0.10 range from the Mincerian literature. It can be due to the fact that the educational variable is endogenous[33]. According to Gennaioli et al. (2013), one way to proxy this variable is by considering the education level of the elderly population, specifically those aged 65 years and older. While this proxy may not be ideal due to potential long-term growth effects, the authors argue that it may suffice since they do not directly intervene in the current production function. Using this proxy yields coefficients that closely align with those obtained using other human capital proxies

[32] For detailed country-level labor shares in 1995, 2007, and 2014, refer to Table A2 in their appendix.

[33] Acemoglu et al. (2014) instrumented the educational variable by the allocation of protestant missionaries in the early twentieth century

	(1)	(2)	(3)	(4)	(5)	(6)
market potential	0.11^{***}	0.09^{***}	0.06^{**}	0.05^{*}	0.07	0.00
	(0.03)	(0.03)	(0.02)	(0.03)	(0.07)	(0.09)
inv. dist. port	0.14^{**}	0.13^{**}	0.14^{**}	0.13^{**}	0.14^{**}	0.13^{**}
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
years education	0.28^{***}	0.28^{***}	0.28^{***}	0.28^{***}	0.28^{***}	0.28^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
population density	−0.00	0.01	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
temperature	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
oil per cap.	0.19^{***}	0.19^{***}	0.19^{***}	0.19^{***}	0.19^{***}	0.19^{***}
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Num. obs.	1,464	1,464	1,464	1,464	1,464	1,464
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Num. groups: code	103	103	103	103	103	103
Adj. R^2 (proj model)	0.43	0.42	0.42	0.42	0.42	0.42
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDPpc_i \tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j \neq i} GDPpc_j \tau_{ij}$ with i and j both in country c.

Table 1.3: Regional Development and Market Potential

and enhances the coefficients associated with the various market potential indexes. However, it has the drawback of significantly reducing the number of observations. The results are presented in Table 1.18 in the appendix.

I compare the results with the ones found by Head and Mayer (2011). They use measures of market potential considering countries, with a similar formulation than $MP^{(h)}$ index. Their index takes into account distance, contiguity, common language, colonial links, and dummies for common membership of a regional trade agreement (RTA), a currency union and WTO membership. Net of the effect of education levels, they find significant coefficients about 0.41 for the year 1995 without country fixed effects, and 0.55 for the period 1965-2005 with country fixed effects. This is sensitively higher than what we find in the similar regressions considering regions in 2005. It suggests that regional development differences within countries are less sensitive to market potential than national development differences worldwide. This observation can be due to the fact that there exist less variation of GDP per capita and market potential within countries than between countries.

Single-country analysis support this statement. The effect of a 1% increase in the regional market potential is found to increase wages by below than 0.2% in general. This results hold for developing countries such as Brazil (Fally et al., 2010), China (Hering and Poncet, 2010; Baum-Snow et al., 2020), Chile (Paredes, 2013), as well as European regions from European analysis

(Niebuhr, 2006; Head and Mayer, 2006; Brakman et al., 2009), and for developed countries such as Germany (Brakman et al., 2004; Kosfeld and Eckey, 2010), Italy (Mion and Naticchioni, 2009), and Spain (Pires, 2006).

Another explanation for smaller coefficients is that I control for more geographic variable than Head and Mayer (2011) do, reducing the potential bias in the estimators. With panel data and controlling for auto-correlation, Boulhol and De Serres (2010) find coefficients between 0.07 and 0.1, as I find in table 1.3.

1.7.2 Robustness

This section presents the robustness checks using Two-Stages-Least-Square estimations on the cross-sectional sample, and the OLS estimation on a panel, but smaller, sample.

2SLS estimations

Two different instruments are used in the Two-Stage-Least Squares estimation. A first one is a centrality index which summarizes the inverse distance to the wealthiest region of each country in the sample, presented in equation 1.7. The second is the foreign market potential, which is not found to have an effect on GDP per capita, as showed by results above.

First, market potential indexes are regressed on the instrument chosen and the regional development covariates, following equation 1.8. Then, the second stage consists in regressing the GDP per capita on the fitted market potential values, as well as the regional development covariates. Column 1 and 2 in table 1.4 show the results of the first and second stage equations, i.e. equations 1.8 and 1.6 respectively. Column 3 and 4 show the results for specifications with the first stage expressed in equation 1.9, using the foreign market potential as instrument.

First stage results show that the higher the centrality of regions in terms of proximity to the wealthier regions from each country, the higher their market potential. More particularly, I find that a region with a 1% higher centrality index than the average region in the same country is predicted to have a 0.08% higher market potential. This result is significant at the 1% significance level. Density is found to be an important determinant of market potential. The larger the population with respect to the regional area, the higher the potential demand and thus, the higher the market potential. Second stage display considerably higher coefficients for market potential indexes than what has been estimated with OLS, suggesting that the effect of regional market potential on development has been under-estimated. Indeed, coefficients are between 0.8 and 0.9 points greater.

In order to check for the instrumentation's relevance, I conduct different tests: (1) endogeneity test, (2) under-identification test[34], and (3) weak identification test. Because it is assumed that the error terms may be correlated at the national level, justifying the use of fixed effects and standard errors clustering at the country level, a first test relies on the Durbin–Wu–Hausman statistic, a second on the Kleibergen-Paap rk LM statistic, and a third on the Kleibergen-Paap rk Wald F statistic. This three tests are robust to heteroskedasticity.

The first test evaluates whether the OLS and the 2SLS estimates are statistically different, i.e. $H_0 : \alpha_1^{(OLS)} \neq \alpha_1^{(2SLS)}$. In other words, it checks for the endogeneity of the market potential variable. The Durbin–Wu–Hausman statistics in column 1 and 2 suggest that the two estimates from the OLS and the IV regressions are statistically different, so that the OLS estimate is

[34]With one instrument, the underidenfication test is equivalent to the IV redundancy test.

	(1)	(2)	(3)	(4)
Second Stage				
temperature	−0.01	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
inv. dist. port	0.18	0.14**	0.14**	0.13**
	(0.14)	(0.09)	(0.06)	(0.06)
oil per cap.	0.24***	0.23***	0.19***	0.19***
	(0.04)	(0.04)	(0.04)	(0.04)
population density	−0.13**	−0.08**	0.00	0.01
	(0.06)	(0.04)	(0.01)	(0.01)
years education	0.26***	0.27***	0.28***	0.28***
	(0.03)	(0.03)	(0.02)	(0.02)
market potential	1.04**	0.87**	0.08	0.00
	(0.51)	(0.42)	(0.08)	(0.09)
Num. obs.	1,464	1,464	1,464	1,464
Country Fixed Effects	Yes	Yes	Yes	Yes
Num. groups: country	103	103	103	103
Adj. R^2 (proj model)	−0.32	0.06	0.43	0.42
Regressor (endogenous variable)	$\ln \text{MP}^{(h)}$	$\ln \text{MP}^{(s)}$	$\ln \text{MP}^{(h)}$	$\ln \text{MP}^{(s)}$
First Stage				
temperature	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)
inv. dist. port	−0.04	−0.00	0.00	−0.08
	(0.14)	(0.11)	(0.15)	(0.13)
oil per cap.	−0.05	−0.06*	−0.02	−0.02
	(0.04)	(0.03)	(0.04)	(0.03)
population density	0.12***	0.09***	0.12***	0.09***
	(0.02)	(0.02)	(0.02)	(0.02)
years education	0.00	−0.01	0.02	0.01
	(0.02)	(0.02)	(0.02)	(0.01)
centrality	0.05**	0.07**		
	(0.03)	(0.03)		
foreign market potential			0.98***	1.02***
			(0.13)	(0.16)
Endogeneity test				
F stat	16.35	13.89	0.23	1.11
	[0.00]	[0.00]	[0.63]	[0.29]
Under-identification test				
Kleibergen-Paap rk LM stat	4.23	5.23	11.12	10.71
	[0.04]	[0.02]	[0.00]	[0.00]
Weak identification test				
Kleibergen-Paap rk Wald F stat	4.07	5.18	55.29	38.77

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. $\text{MP}^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $\text{MP}^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. The variables $\text{centrality}^{(h)}$ and $\text{FMP}^{(h)}$ ($\text{centrality}^{(s)}$ and $\text{FMP}^{(s)}$ respectively), are used as instruments for $\text{MP}^{(h)}$ ($\text{MP}^{(s)}$ respectively).

Table 1.4: IV results

inconsistent. However, the second and third IV tests reveal statistics lower than 10[35]. Thus, we can not reject that the instrument is weak, neither that it is redundant. In other words, the instrument seems not to be relevant and the estimated coefficient may be biased.

Since the OLS coefficients for the effect of foreign market potential indexes on regional GDP per capita are not significant, I choose to use them as instruments[36]. The first stage shows highly significant elasticity coefficients around unity, with Kleibergen-Paap statistics higher than 10 and 16.38 for the under-identification and weak identification tests respectively. However, the endogeneity test reveals that the effect of the fitted market potential is not different from the OLS specification, with non-significant IV coefficients. This result suggests to believe in the coefficients about 0.1 estimated in column 1 and 2 from table 1.3. However, 2SLS coefficients are found to be not significant. It could come from the fact that market potential impact regional inequality differently depending on the income group of the country, as explained in section 1.7.3.

Head and Mayer (2006), who estimate market potential indexes for European regions with a similar method, find coefficients about 0.12 in their OLS estimation, and about 0.07 in their IV estimation. These coefficients are similar to those found from the OLS specification - see Table 1.3.

Panel estimations

To test the robustness of the cross-sectional results, we estimate regression 1.6 with the panel data available for the years 1995, 2000, 2005 and 2010. The panel sample comprises 1,064 subnational regions in 72 countries, although the dataset is unbalanced. On average, each region appears in the sample for 2.9 years, with a minimum of one year and a maximum of four years. However, due to a lack of information on education for some regions during the years 2005 and 2010, a significant number of observations will need to be removed from the estimations. To address this issue, I merge the panel dataset with the 2005 cross-sectional dataset to complete missing observations on education for that year. This will help to improve the completeness of the sample and ensure that the estimations are based on a more representative dataset. First, cross-section regressions with country fixed effets are estimated for each year of the panel sample. Then, panel estimations are performed with fixed effects at the country-year level, ζ_{ct}, to control for country-time varying unobservables. Results are displayed in table 1.21 in the appendix.

Elasticity coefficients to market potential from the different cross-sectional estimations are similar than those found in table 1.3 - about 0.1 - for the years 1995, 2000 and 2005. Foreign market potential elasticity coefficient is not significant in 2005, as found in table 1.3, but is significantly equal to 0.1 in 1995 and 0.2 in 2000. However, results for the year 2010 show unsignificant coefficients for market potential indices computed with the haversine distance, while indices computed with the shipment function are found to have a significant negative effect on income per capita. Results may be biased as a result of a lack of observations, i.e. only 20 countries are considered.

Results for the complete panel estimations support that a region with a market potential that is 1% higher than the average region within country-year groups is predicted to have a

[35]The Kleibergen-Paap rk Wald F statistic from the weak identification test is compared to the Stock-Yogo weak ID test critical values for a 10% maximal IV size, which is about 16.38. Hence, the null hypothesis is not rejected.

[36]Note that the foreign market potential is redundant when adding the variable in equation 1.6 to be estimated.

0.1% higher income per capita. Coefficients display a higher significance level when excluding the year 2010 (see Table 1.21 in the appendix). It is also found that proximity to foreign markets do matter even more significantly than the non-local market potential, while the reverse has been found in cross-section.

Table 1.22 in the appendix presents the results for the investigation of heterogenous effect of market potential regarding the core-periphery structure of regions within countries. Results still show that peripheral regions robustly display elasticity coefficients lower by 0.01 to 0.02 percentage points lower than core regions.

1.7.3 The Core and Periphery Divide

Baseline regressions

The heterogeneous effect of market potential on regional development is investigated with respect to the income position of regions within countries, classified into three different groups: the *core*, the *semi-periphery* and the *periphery*. This specification is motivated by the Krugman (1991)'s Core-Periphery model which explains how the economic activity can concentrates in specific regions, leaving the periphery behind. The lower the transportation and trade costs to a wealthy region, the higher the agglomeration forces towards this wealthy regions. The poorer a market, the higher the agglomeration forces towards wealthy regions. Following this statement, we can wonder if market potential can have a negative effect on development for peripheral regions.

Table 1.5 shows the results, with the interacted effect of market potential with the regional group dummies. The coefficient of market potential gives the effect of market potential on core regions - the wealthiest in terms of GDP. A 1% increase in the market potential in core

	(1)	(2)	(3)	(4)	(5)	(6)
market potential	0.12^{***}	0.10^{***}	0.08^{***}	0.07^{***}	0.09	-0.00
	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)	(0.08)
market potential $\times \mathbb{1}_{\text{semi-periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times \mathbb{1}_{\text{periphery}}$	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Num. obs.	1460	1460	1460	1460	1460	1460
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Num. groups: code	101	101	101	101	101	101
Adj. R^2 (proj model)	0.47	0.47	0.47	0.47	0.46	0.46
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDP_i \tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j \neq i} GDP_j \tau_{ij}$ with i and j both in country c.

Table 1.5: Regional Development and Market Potential - Core and Periphery

	(1)	(2)	(3)	(4)	(5)	(6)
High income countries						
market potential	0.04	0.04	0.02	0.03	0.14	0.09
	(0.03)	(0.03)	(0.02)	(0.03)	(0.10)	(0.09)
market potential $\times$ $\mathbb{1}_{\text{semi-periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times$ $\mathbb{1}_{\text{periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Num. obs.	441	441	441	441	441	441
Num. groups: code	29	29	29	29	29	29
Adj. R^2 (proj model)	0.42	0.42	0.42	0.42	0.43	0.42
Upper-middle income countries						
market potential	0.16^{***}	0.13^{**}	0.12^{**}	0.10	0.14	-0.06
	(0.05)	(0.06)	(0.05)	(0.06)	(0.13)	(0.17)
market potential $\times$ $\mathbb{1}_{\text{semi-periphery}}$	-0.01	-0.01	-0.01^{*}	-0.01^{*}	-0.01^{*}	-0.01^{*}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times$ $\mathbb{1}_{\text{periphery}}$	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Num. obs.	519	519	519	519	519	519
Num. groups: code	29	29	29	29	29	29
Adj. R^2 (proj model)	0.52	0.51	0.51	0.51	0.50	0.50
Lower-middle income countries						
market potential	0.21^{**}	0.25^{**}	0.14^{***}	0.18^{***}	-0.00	0.05
	(0.09)	(0.10)	(0.04)	(0.06)	(0.15)	(0.23)
market potential $\times$ $\mathbb{1}_{\text{semi-periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{**}	-0.01^{**}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times$ $\mathbb{1}_{\text{periphery}}$	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Num. obs.	368	368	368	368	368	368
Num. groups: code	29	29	29	29	29	29
Adj. R^2 (proj model)	0.51	0.51	0.50	0.50	0.49	0.49
Low income countries						
market potential	0.27	-0.01	0.14	-0.09	1.21	0.30
	(0.52)	(0.65)	(0.50)	(0.62)	(1.04)	(1.03)
market potential $\times$ $\mathbb{1}_{\text{semi-periphery}}$	-0.01^{*}	-0.01^{*}	-0.01^{*}	-0.01^{*}	-0.01^{*}	-0.01^{*}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
market potential $\times$ $\mathbb{1}_{\text{periphery}}$	-0.02^{**}	-0.02^{**}	-0.02^{**}	-0.02^{**}	-0.02^{**}	-0.02^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Num. obs.	132	132	132	132	132	132
Num. groups: code	14	14	14	14	14	14
Adj. R^2 (proj model)	0.55	0.54	0.54	0.54	0.56	0.54
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDP_i\tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j\neq i} GDP_j\tau_{ij}$ with i and j both in country c.

Table 1.6: Regional Development and Market Potential - by countries' income group (2005)

regions is expected to increase their GDP per capita by 0.1%. If the region is categorized as a periphery region, the expected effect of market potential on regional development is 0.02 percentage points lower, but still positive. The same result holds for non-local market potential. While proximity to wealthy foreign regions is found to have no effect on core regions, (semi-)peripheral regions are found to have an elasticity coefficient significantly lower by (0.01) 0.02 percentage points. This result suggest that the (semi-)periphery could suffer from too much proximity to foreign markets. A 1% increase in the foreign market potential in peripheral regions can be associated to a decrease by 0.02% in the GDP per capita.

The elasticity coefficient should theoretically equate to $1/(\beta\sigma)$ as mentioned earlier, where β represents the income labor share and σ signifies the elasticity of substitution between varieties. The observed lower coefficient linked to the periphery, as compared to the coefficient associated with the core, could arise from either a higher income labor share or a greater elasticity of substitution between varieties. The former explanation is plausible in regions where agriculture serves as the predominant local economic activity. In such areas, which often contend with elevated transportation and trade costs, market potential may hold less relevance. A higher elasticity of substitution between varieties might depend on the quality of the goods produced and the industry (Fontagné et al., 2022).

Countries' income group regressions

Table 1.6 displays the estimated coefficients of the effect of market potential and its proxies on regional development by countries' income group following the World Bank classification: *High income*, *Upper middle income*, *Lower middle income* and *Low income*. Results show an heterogeneous effect of market potential. Especially, the split of the sample do not give significant estimate of the effect of market potential for high income and low income countries. Moreover, differences in market potential explains higher regional development differences in lower middle income countries than in upper middle countries. The former display elasticity coefficients about 0.21, 0.25 for $\text{MP}^{(s)}$, while the latter about is 0.16, 0.13 for $\text{MP}^{(s)}$. This result is in line with Boulhol and De Serres (2010) who find that, in the sample of Redding and Venables (2004), higher income OECD countries display a lower coefficient elasticity of GDP per capita to market access than for the whole sample of OECD countries, which is also lower than coefficients for the whole sample. Hence, they show that pooling developed and developing countries in country-level analysis lead to upward biased estimators (Redding and Venables, 2004; Head and Mayer, 2011).

The current findings align with the research conducted by Brülhart et al. (2020), which delves into agglomeration and dispersion effects on economic activity based on a country's developmental stage. Their model features one central and two peripheral locations, exploring the influence of decreasing trade costs on employment location. Initially, when trade costs are exceedingly high, a location's connectivity offers no discernible advantage, resulting in an even distribution of employment across all three locations. However, as trade costs decrease, the central location gains employment at the expense of the peripheries due to enhanced access. Subsequently, as trade costs continue to diminish, the central location loses its proximity advantage, rendering peripheral locations attractive once more due to reduced congestion and improved market access.

Brülhart et al. (2020)'s model and results provide insight into the absence of significant market potential influence on regional development disparities within high-income countries, often characterized by high connectivity and low transportation cost. In such contexts, the

periphery is expected to converge gradually toward the level of development seen in core regions. Conversely, market potential plays a substantial role in driving development disparities within middle-income economies, where trade costs remain substantial, the economic activity tends to concentrate in areas with high market potential, often at the expense of the periphery. Furthermore, these insights help clarify the lack of significant outcomes observed in low-income economies, where transportation infrastructure remains undeveloped, and the proximity advantage has little influence on the location of employment and GDP per capita, both of which may be widely dispersed across regions.

Insignificant results may also stem from the dataset containing regions that are larger in size than what we would ideally prefer for examining economic activity concentration and disparities. In the regression results from table 1.5, we note that there are 1,464 observations included in the estimation, spanning across 103 countries, which leads to an average of roughly 14 regions per country — a relatively small number for analysis. In particular, there are an average of 15 regions per high-income country, 18 per upper-middle income countries, 13 per lower-middle income countries and 9 per low-income countries. Additionally, Table 1.16 reveals limited variation in market potential values between core and peripheral regions within countries, with the coefficient of determination ranging from approximately 5% to 7% for market potential, 3% for non-local market potential, and 0% to 1% for foreign market potential.

Summing up a multitude of values in the market potential indexes poses a challenge in capturing the regional disparities that distinguish core from periphery regions. This challenge is exacerbated by the limited differentiation in market potential values between these regions within countries. To delve deeper into the influence of proximity to core markets and tackle the constraints imposed by market potential indexes, I incorporate measures of centrality for core regions into the regression analysis. This approach seeks to introduce additional variation among regions within countries, based on their proximity to significant demand pools. The following section presents the results of this effort to address the aforementioned limitation.

Centrality to domestic and foreign cores

Results above have suggested that proximity to foreign markets is not necessarily beneficial for development, especially for the (semi-)periphery. It is natural to think that, within countries, regions with the strongest foreign market potential are located at the national borders. Hence, it is possible to relate this result to the trade border literature introduced by McCallum (1995). International trade is found to be more costly than intra-national trade among regions within a country. In addition, while the discussion is still vivid, there is a consensus that international trade increases income (Frankel and Romer, 1999; Anderson et al., 2020). As a result, if international borders decrease trade between regions, and if trade leads to income growth, regions at the border may display lower income levels than the national average since their location places them further to central national regions and closer to foreign regions. This statement is supported by Adam et al. (2023) and Bonadio et al. (2023), where they find that subnational regions at the international borders have lower income per capita levels and night lights intensity, while trade agreements at borders compensate this negative border effect and ultimately reduce regional disparities. On the hand, Brülhart (2006) finds that both peripheral and core regions of countries experienced growth in the context of European Union

integration.[37]

Another explanation for a potential negative impact of a close proximity to foreign markets is spatial sorting of firms within countries and market selection. Combes et al. (2012a) highlight how denser cities foster competition, allowing only the most productive firms to survive. As a result, firms are spatially sorted based on their productivity, with peripheral regions typically accommodating the least productive firms. On another hand, Melitz and Ottaviano (2008) argues that easier international trade increases competition due to good access to foreign goods, leading to the demise of less productive firms and an overall higher productivity.

Among core domestic regions, firms selection effect may result in higher levels of productivity and higher development. In contrast, peripheral regions, which tend to host less productive firms, can be hurt from proximity to core foreign markets if firms cannot face the international competition. Based on this assumption, the impact of proximity to core markets with a free trade agreement is expected to be negative for peripheral regions, and higher in magnitude than proximity to core markets without a free trade agreement.

In order to expand upon the previous assertion, I examine the heterogenous effects of centrality to core domestic and foreign markets, for each group of regions, *the core*, *the semi-periphery* and *the periphery*. I regress regional per capita income on development covariates with fixed effects at the country level, as well as at the cluster level, and include the centrality indexes as independent variables. To account for the impact of free trade agreements and for comparability with the results found by Adam et al. (2023) and Bonadio et al. (2023), I also differentiate between centrality to foreign cores with and without such agreements. Results are presented in table 1.7. Table 1.23 in the Appendix includes the market potential index as a control variable, to get ceteris paribus effect of overall market access - results are similar. Table 1.8 examines the relationships within four distinct country samples, categorized based on their income levels.

Table 1.7 shows that overall centrality to cores have no effect on regional development. However, when decomposing the index of centrality to cores into its domestic and foreign components, it is found that centrality to domestic cores has a positive impact on regional development, while centrality to foreign cores has a negative impact, especially for peripheral regions (see columns 3 and 4). The negative impact is primarily driven by proximity to foreign cores that have no free trade agreement with the region's country (see columns 5 and 6). It tends to approve the hypothesis that periphery regions closer to the border face higher trade costs since they are far from domestic markets in general, and close to foreign countries that apply tariffs. Proximity to foreign core regions with free trade agreements does not significantly contribute to explaining regional disparities within countries. This result prompts questions regarding the validity of the hypothesis suggesting that proximity to foreign core markets has detrimental effects on peripheral regions due to firm selection. However, I refrain from making definitive conclusions on this matter since my available data does not allow for a conclusive assessment.

Table 1.23 in the appendix presents results with market potential control. Including the market potential index in the regression reveals a negative effect of centrality to domestic cores for core regions, although this effect is not statistically significant (columns 7 and 8). This finding underscores the spatial division that arises from economic agglomeration, where cores are located at a certain distance from each other, and peripheral regions fill the space between

[37] In particular, Brülhart (2006) finds that countries' peripheral regions grew in terms of manufacturing employment, while core regions grew in terms of service employment.

them. Nevertheless, the positive and significant effect of market potential still supports the idea that regions closer to domestic cores tend to be wealthier. On the other hand, centrality to foreign cores remains a significant predictor of regional development, even with the market potential control. The closer regions are to foreign cores, the poorer they tend to be, with this relationship particularly pronounced for peripheral regions (columns 7 and 8), everything else being equal.

Table 1.8 displays the regression estimation derived from column 6 of Table 1.7, by splitting the sample into the four countries groups, categorized based on their income levels. The results consistently demonstrate a positive effect of proximity to domestic core markets, with the exception being low-income countries. This discovery may appear to contradict the findings of Baum-Snow et al. (2020), who observed that connectivity to national core markets in China had an adverse impact on regional income and population growth in the hinterlands. However, if both regional income and population size decrease, it is possible for regional income per capita to remain stable or even improve, maintaining or increasing the level of development. It is worth highlighting that China is categorized as an upper-middle-income country within the dataset, and this specific group of countries displays the most notable positive coefficient for peripheral regions.

Proximity to foreign cores without a trade agreement consistently exhibits a negative impact, with more pronounced effects observed in peripheral regions. This negative effect is highly significant in both high- and low-income countries. In contrast, proximity to foreign cores with a trade agreement is found to have a positive and statistically significant impact in lower-middle-income countries, especially for core and semi-periphery regions, which advocates for the positive impact of trade liberalization. However, the coefficient for the periphery is not significant.

To gain further insights into the impact of free trade agreements, I employ panel data to examine whether the act of signing a free trade agreement contributes to regional development, with a specific emphasis on the periphery, which is the primary subject of concern. Table 1.24 in the appendix shows the results. Results show no effect in the change of centrality to foreign cores resulting from the signing of an agreement, for any of the core and the periphery.

Further research is needed to fully understand the dynamics of proximity to foreign markets, market selection, and regional development. Future studies should employ micro-founded analyses to explore the specific mechanisms of firms' selection, consider the heterogeneous impacts of trade on different regions, and account for local economic factors. By delving deeper into these aspects, researchers can validate and refine the assumptions made in this study, ultimately providing a more comprehensive understanding on challenges face by peripheral areas.

1.8 Conclusion

The main objective of this paper is to explore the key factors contributing to regional development disparities within countries and investigate the role played by proximity to markets on the core-periphery divide. This paper contributes to provide a falsification test to the international trade wage equation developed by Fujita et al. (1999) at the regional level with a worldwide scope. By analyzing these factors across a wide range of regions and countries, this research aims to provide a more comprehensive understanding of the international trade wage equation and its implications for regional disparities.

I use an extended regional dataset which includes more than 1,500 subnational regions, from 107 different countries, out of the 195 recognized around the world. This dataset contains information on regional economic activity, geography and education, to control the estimations for these regional development covariates. Hence, market potential indexes are built at the regional level, following a gravity-based approach. One of the contributions of the paper lies in the computation of a more accurate measure of distance between markets, which considers land masses and ocean areas in the market potential distance function. I compute the inland shortest path between regions and ports, as well as the overseas shortest path between ports.

Results provide evidence that access to markets has a significant impact on regional development for both specification of market potential presented. In particular, proximity to demand is associated to a more intense economic activity and a higher income per capita. In other words, regions surrounded by rich markets may be richer than regions with poor neighbours, due to lower trade costs as stated by the wage equation, but also to positive externalities. This statement comes the fact that the market potential index computed on the basis of the haversine distance function seems equally robust than the one constructed with the shipment distance function. It suggests that other things than goods may flow from a region to another, which does not rely on port connection, such as knowledge and people.

Connecting regions between each other seems to be an important challenge to tackle in order to develop and increase living conditions everywhere. However, heterogeneous effects exist conditional on the national income group, as well as on the regional core-periphery structure within countries. On one hand, the effect of market potential is found to be particularly strong in both upper- and lower-middle-income countries. On the other hand, (semi-)peripheral regions are found to exhibit lower elasticity coefficients to market potential compared to core regions. Additionally, results show that being in close proximity to foreign core markets can be detrimental to regions located in the (semi-)periphery. Peripheral regions tend to have lower per capita income and lower market potential compared to core regions. Additionally, peripheral regions experience slower growth rates over time, which further widens the income gap between them and core regions. Moreover, due to their lower development elasticity coefficient to market potential, peripheral regions may be less responsive to the growth of surrounding markets, which could leave them even more disadvantaged in the long run.

Further investigation highlights that the impact of proximity to core markets varies depending on whether they are domestic or foreign, as well as whether a region's country has a free trade agreement with others. Proximity to domestic cores is associated with positive effects on regional development, while proximity to foreign cores without a trade agreement has a negative impact for peripheral regions. Interestingly, the effect of proximity to foreign cores with a free trade agreement is not significantly different from zero, suggesting that the presence of a free trade agreement appears to mitigate the negative impact of close proximity to foreign core markets, or at least reduce regional disparities. Indeed, the analysis using panel data indicates that the act of signing a trade agreement, leading to an increase in centrality to foreign cores with FTA, is associated with a rise in GDP per capita, although the statistical significance is modest. These results imply that the existence of a free trade agreement shows potential in fostering regional development, including in peripheral areas, as also evidenced in the literature (Brülhart et al., 2004; Adam et al., 2023; Bonadio et al., 2023).

These findings underscore the importance of understanding the unique challenges faced

by peripheral regions in achieving economic development and suggest the need for policies that are tailored to the specific circumstances of these regions. One policy suggestion that arises from the study is to enhance connectivity between peripheral regions and core domestic regions in order to reduce transportation costs and increase access to domestic demand. Nonetheless, it's worth noting that this policy might not be the most suitable approach for low-income countries, as there's a potential risk of regional economic activity and population relocating from the periphery to existing core regions. Additionally, policies aiming at improving trade relations and establishing free trade agreements with foreign countries may be beneficial to peripheral regions.

Appendix to chapter 1

1.A Appendix : Regional Trade Gravity Equation and Trade Elasticity

1.A.1 Regional Trade Gravity Equation

To express and compute regional market potential indexes, I estimate the gravity equation in order to use the coefficients of trade elasticity to trade costs as explained in section 1.3. The gravity model states that the force of attraction between two entities depends positively on their respective mass and negatively on the distance between them - the distance playing as a resistance force. Applied to trade, it states that the wealthier and the closer two markets are, the more intense their bilateral trade activity. The underlying mechanisms follow a supply and demand analysis. The larger the exporter's production, the more it is going to sell. The larger the importer's demand, the more it is going to expend. If both markets are close, transportation costs from one to another are low. Thus, the two markets are likely to fulfill their need: to trade between each other.

To estimate the gravity equation, I use transnational trade panel data and estimate the expected trade flows between regions of each pair of countries. I assume that the bilateral trade flows between two regions of different countries is proportional to the share of their regional Gross Domestic Product (GDP) in the overall national GDP. Hence, the wealthier the region, the more it participates to the wealth of its country, and the more intense its international activity. Hence, bilateral trade flows between regions are proxied by:

$$\tilde{\text{TF}}_{ij} = \text{TF}_{c_i c_j} \times \frac{\text{GDP}_i}{\text{GDP}_{c_i}} \times \frac{\text{GDP}_j}{\text{GDP}_{c_j}} \tag{1.19}$$

with $\tilde{\text{TF}}_{ij}$ the expected bilateral trade flows between region i in country c_i and region j in country c_j, with $c_i \neq c_j$, $\text{TF}_{c_i c_j}$ the observed value of bilateral trade flows observed between countries c_i and c_j in the data, GDP_i and GDP_{c_i} the Gross Domestic Product of region i and its country c_i respectively. The closer the ratio $\text{GDP}_i/\text{GDP}_{c_i}$ to unity, the more country c_i's economic activity relies on the one of region i. The same applies for the ratio $\text{GDP}_j/\text{GDP}_{c_j}$.

Then, the gravity equation is expressed as follows:

$$\tilde{\text{TF}}_{ij} = A \times \text{Y}_i^{\alpha_1} \times \text{Y}_j^{\alpha_2} \times \tau_{ij} \tag{1.20}$$

with A a constant, $\tilde{\text{TF}}_{ij}$ the expected total bilateral trade flows between regions i and j, Y_i and Y_j their respective mass, and τ_{ij} the resistance term -or proxy of trade costs- between them. The latter variable has been expressed in two manners, as showed in equations 1.3 and 1.5.

The gravity equation is estimated in cross-section as follows:

$$\begin{aligned}\ln \tilde{\text{TF}}_{ij} &= \beta_1 \ln \text{dist}_{ij}^{(\text{haversine})} + \beta_2 \mathbb{1}_{\text{language}_{ij}} + \beta_3 \mathbb{1}_{\text{contiguous}_{ij}} + \beta_4 \mathbb{1}_{\text{colony}_{ij}} \\ &\quad + \beta_5 \mathbb{1}_{\text{rta}_{ij}} + \beta_6 \mathbb{1}_{\text{currency}_{ij}} + \delta_i + \delta_j + \epsilon_{ij} \end{aligned} \tag{1.21}$$

$$\begin{aligned}\ln \tilde{\text{TF}}_{ij} &= \gamma_1 \ln \text{dist}_{io}^{(\text{land, from exporter to origin port})} \mathbb{1}_{\text{maritime route}} \\ &\quad + \gamma_2 \ln \text{dist}_{od}^{(\text{sea, between ports})} \mathbb{1}_{\text{maritime route}} \\ &\quad + \gamma_3 \ln \text{dist}_{dj}^{(\text{land, from destination port to importer})} \mathbb{1}_{\text{maritime route}} \\ &\quad + \gamma_4 \ln \text{dist}_{ij}^{(\text{land})} (1 - \mathbb{1}_{\text{maritime route}}) \\ &\quad + \beta_2 \mathbb{1}_{\text{language}_{ij}} + \beta_3 \mathbb{1}_{\text{contiguous}_{ij}} + \beta_4 \mathbb{1}_{\text{colony}_{ij}} + \beta_5 \mathbb{1}_{\text{rta}_{ij}} + \beta_6 \mathbb{1}_{\text{currency}_{ij}} \\ &\quad + \delta_i + \delta_j + \epsilon_{ij} \end{aligned} \tag{1.22}$$

with $\text{dist}_{ij}^{(\text{land})} = \kappa_{ij}$, $\text{dist}_{io}^{(\text{land, from exporter to origin port})} = \kappa_{io}$, $\text{dist}_{dj}^{(\text{land, from destination port to importer})} = \kappa_{od}$ and $\text{dist}_{ij}^{(\text{sea, between ports})} = \kappa_{od}$ from equation 1.5. The dummy $\mathbb{1}_{\text{maritime route}}$ equals 1 if the condition expressed in equation 1.4 is respected, i.e. if it is more convenient for region i to send its goods overseas to region j, and 0 otherwise. $\tilde{\text{TF}}_{ij}$ is the expected total bilateral trade flows between regions i and j, in countries c_i and c_j respectively, with $c_i \neq c_j$. Since I estimate the effect of distance on trade flows which is time invariant, I cannot use pairs fixed effects to reduce endogeneity problems. However, I include fixed effects for both regions of the pair, δ_i and δ_j, in order to control for the unobservable and unit-specific confounders in 2005. Therefore, it controls for supply, demand and multilateral resistance terms of each region.

The trade literature has widely estimated coefficients α_1, α_2 and β_1 in cross-section and repeated cross-section estimations at the national level. Coefficients α_1 and α_2 are generally found to equal 1, and β_1 to -1. This result is stable across different periods of time and samples of countries. Coefficients β_2, β_3, β_4, β_5 and β_6 estimates are supposed to be positive, as they play as trade facilitators. Next subsection presents the results of the estimated coefficients of trade elasticity to trade resistance and facilitator terms at the regional level using expected bilateral trade flows and compares the results with those found in the literature.

1.A.2 Regional Trade Costs Parameters Estimations

Table 1.9 shows the estimated coefficients from equations 1.21 and 1.22 in columns 1 and 2 respectively. Without surprise, the estimator of trade elasticity to physical distance is negative regardless of the specification of distance. Trade elasticity to distance, as measured as haversine, is about -1.18 (column 1), meaning that a 1% increase in the great-circle distance in kilometers between two regions of different countries decreases the total bilateral trade flows by 1.18% between them.

Distance as measured by the shortest path in kilometers between regions i and j considering land and ocean areas also give negative coefficients of trade elasticity. When it is considered to be convenient for the shipment to be conducted on maritime routes between regions i and j, trade elasticity to distance from regions to their closest port is estimated to be about $[-0.07; -0.06]$, and between ports about -0.96. Otherwise, for those pairs that are considered to ship goods by land routes only, the trade elasticity coefficient is -1.

As expected, contiguity, common language, colonial ties and being part of a regional trade agreement play as trade facilitator forces. Indeed, the corresponding coefficients are found

to be significantly positive. Trade elasticity to contiguity and colonial ties are found to be greater than 1. Unexpectedly, the dummy of common currency gives a negative and significant coefficient. It suggests that two regions belonging to two different countries tend to trade less between each other if their countries share a common currency, whereas it is supposed to be a trade facilitator as well.

Section 1.A.3 in the appendix presents the results obtained from estimations undergone at the country level, compares the coefficients with those found in the literature and presents the statistical tests conducted. Elasticity coefficients estimated here are found to be similar to estimates from the country level analysis, as well as to coefficients usually obtained from gravity equation estimations in the literature (see table 1.11), except for the effect of sharing a common currency.

For the computation of the market potential indexes, I am going to use estimators of distance, and the estimates of trade facilitator found in column 2 table 1.9, except for the effect of currency which may be badly estimated. Instead, I use the average coefficient of 0.79 estimated in the literature (see column 1 of table 1.11). Since I express the market potential of subnational regions, I consider regions belonging to same countries as well. However, the worldwide coverage data I use for the gravity equation estimations do not contain intra-national trade flows. The coefficient β_7, which represents the border effect has been estimated to equal 1.96 on average, such as displayed by the coefficients of the variable home_{ij} from the meta-analysis of Head and Mayer (2014) (see column 1 of table 1.11).

Therefore, the vectors of estimators is the following: $\left[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4\right] = [-1.18, 0.66, 1.12, 1.37, 0.47, 0.79, 1.96, -0.07, -0.96, -0.06, -1.00]$ From these estimates coefficients, market potential indexes are computed following equation 1.1.

1.A.3 Robustness - Trade Elasticity Estimates Comparison

For robustness, regression 1.21 is estimated at the country level with the observed trade flows between countries as dependant variable. Results are displayed in table 1.10. The different columns display the results for different samples of countries. Column 1 includes countries both present in BACI and Gennaioli datasets for the year 2005. Column 2 includes all countries present in BACI dataset for the year 2005. Column 3 and 4 follow columns 1 and 2 respectively, but including all the years between 1996 and 2010.

Coefficients found with countries as level of geographic unit are very similar. Regarding the effects of having a common currency is still found to be negative, but statistically insignificant for specifications in columns 2 to 4. Hence, the negative coefficient displayed may result from an endogeneity problem and/or a sample selection bias. Further investigation shows that the negative sign becomes positive when adding pair fixed effects. The coefficient takes value around 0.17 with the sample of regression in column (3), significant at the 1% level, and 0.16 with the sample of regression in column (4), significant at the 5% level.

The significance tests conducted over the coefficients in tables 1.9 and 1.10 examined whether coefficients where statistically different from zero in order to assess the effect of the presented variable on trade. In other words, the test conducted can be written as: $H_0^{(1)} : \beta = 0$, with $H_1^{(1)} : \beta \neq 0$ the alternative, with $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \gamma_1, \gamma_2, \gamma_3, \gamma_4]$ the set of coefficients to be estimated.

I conduct other statistical tests here. The second test evaluates whether the estimated coefficients in the expected regional trade flows are equal to the ones found from the esti-

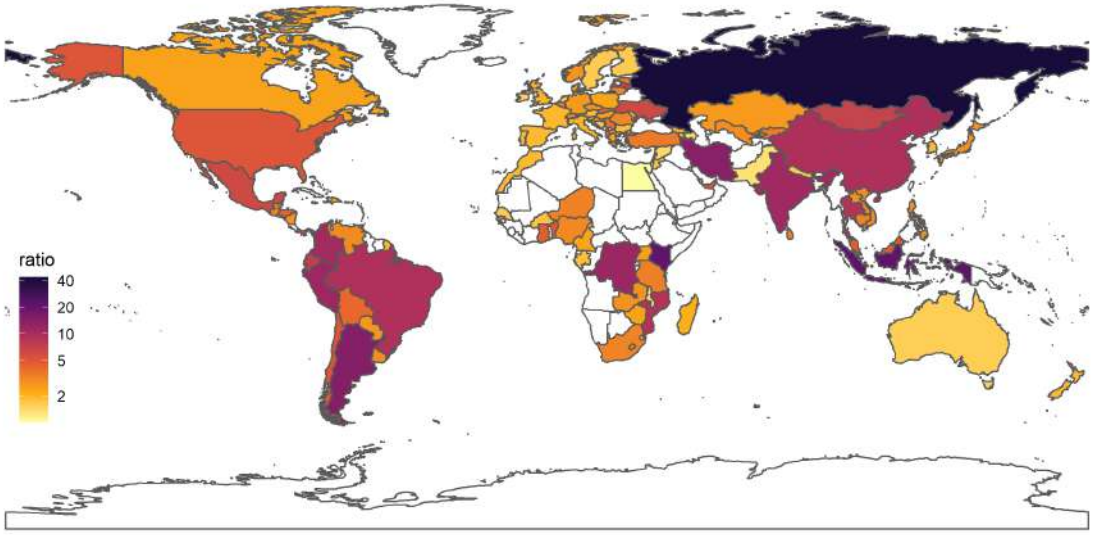
mations conducted at the national level. The corresponding hypothesis can be written as: $H_0^{(2)} : \beta^{(1)} = \hat{\beta}_{C1}$, and $H_1^{(2)} : \beta^{(1)} \neq \hat{\beta}_{C1}$ the alternative, with $\hat{\beta}_{C1}$ the set of estimated coefficients table 1.10 column 1. The same test is undertaken for the estimates $\beta^{(2)}$ in table 1.10 column 2.

The third and fourth tests evaluate whether the estimated coefficients from national estimations are equal to the ones found from the meta-analysis of all gravity papers and of structural gravity papers used by Head and Mayer (2014). The hypothesis can respectively be written as: $H_0^{(3)} : \beta_{C1} = \hat{\beta}_{HM1}$, with $H_1^{(3)} : \beta_{C1} \neq \hat{\beta}_{HM1}$ the alternative, and $H_0^{(4)} : \beta_{C1} = \hat{\beta}_{HM2}$, with $H_1^{(4)} : \beta_{C1} \neq \hat{\beta}_{HM2}$ the alternative. The sets of estimated coefficients $\hat{\beta}_{HM1}$ and $\hat{\beta}_{HM2}$ are transcribed in columns 1 and 2 of table 1.11 respectively. Finally, the same hypothesis are tested with $\beta^{(1)}$ and $\beta^{(2)}$.

Table 1.12 presents the p-values of the different statistical tests. Columns 1 and 2 show that all variables give elasticity coefficients significantly different from the regional to national gravity equations, the former using expected inter-regional trade flows and the latter using observed international trade flows. This is true except for the haversine distance and the dummy of common currency. Despite the significant difference between estimates, they are similar.

Then, coefficients from the country-level-estimations in table 1.10 are compared to those from the literature displayed in table 1.11. The haversine distance is found to be statistically different from -0.93 but not from -1.10. Among the trade facilitator variables, contiguity, colonial ties and regional trade agreement are found to not be statistically different from the average coefficients estimated in the literature. However, the effect of sharing a currency is unexpectedly found to be negative, and especially to be equal to the opposite of the ones found in the literature.

1.B Appendix: Figures



Note: The relationship between the two measures of inequality, the Gini index (see Figure 1.1) and the GDP ratio, display a significant correlation of approximately 0.71. However, the relationship is non-linear, shifting the ranking of inequality. Russia has the greatest level of inequality, with its richest region, Tyumen, being 43 times wealthier than the poorest region, Republic of Ingushetia. This disparity is not surprising, as the country's economic activity is concentrated in a few regions, while others have lower levels due to geography and climate. Indonesia and Kenya follow with a ratio of 20. Argentina, Iran, the Democratic Republic of the Congo, Peru, India, Colombia, and China follow with ratios decreasing towards 9. The least unequal countries have ratios between 1 and 1.5, except for France where the wealthiest region is twice as wealthy as the poorest in terms of GDP per capita. While both measures of inequality are imperfect, they provide insight into existing regional disparities within countries.

Figure 1.3: Ratio highest/lowest regional income per capita in 2005

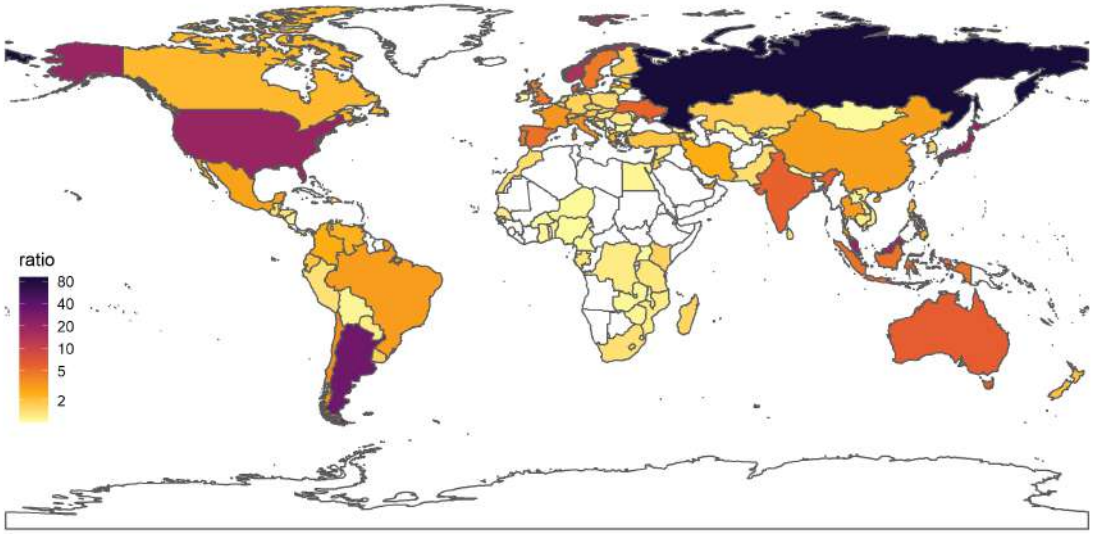


Figure 1.4: Ratio highest/lowest regional market potential $MP^{(s)}$ in 2005

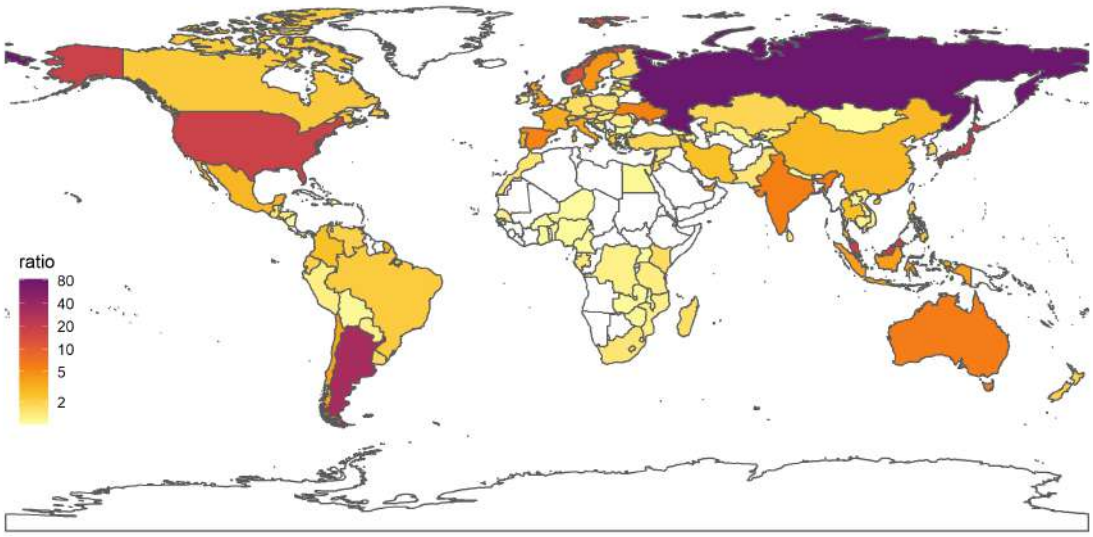


Figure 1.5: Ratio highest/lowest regional non-local market potential $NLMP^{(s)}$ in 2005

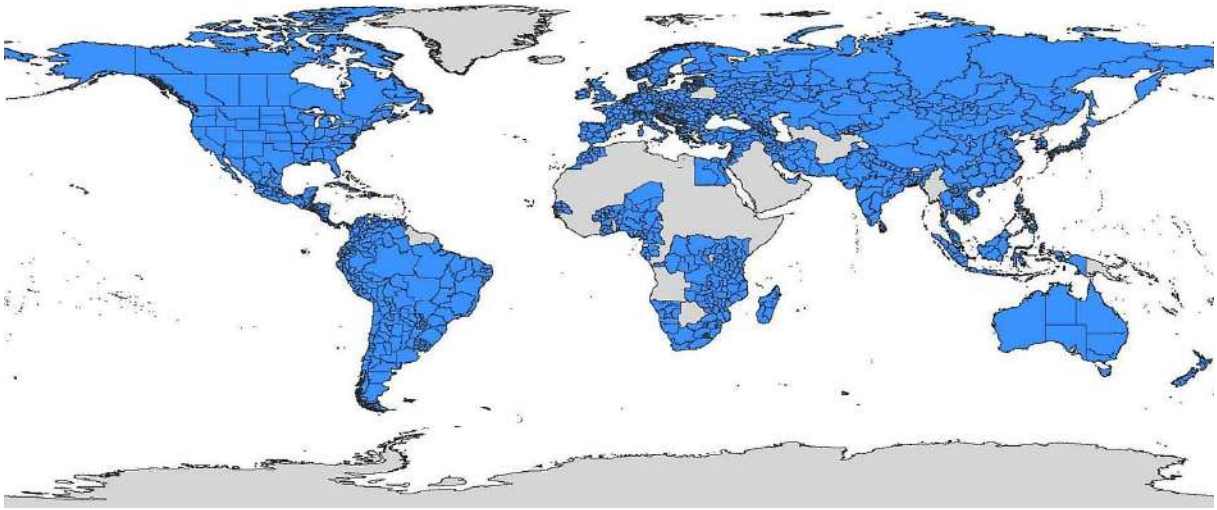


Figure 1.6: Gennaioli et al. (2013) regional dataset

1.C Appendix: Tables

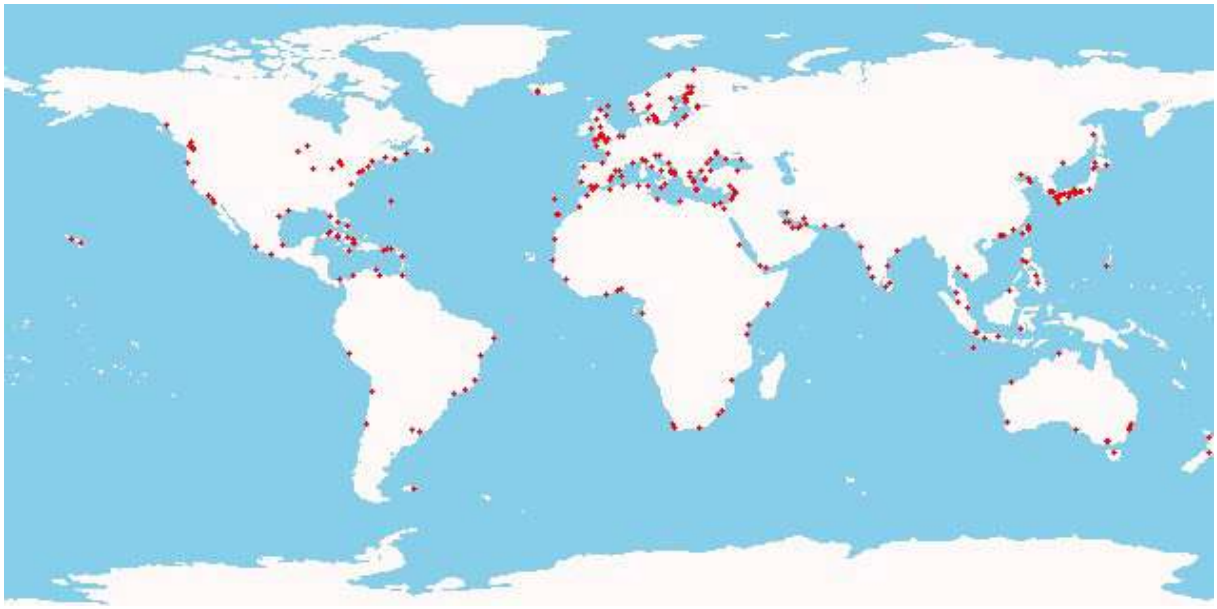


Figure 1.7: World ports selected as the closest ports to each region in the sample

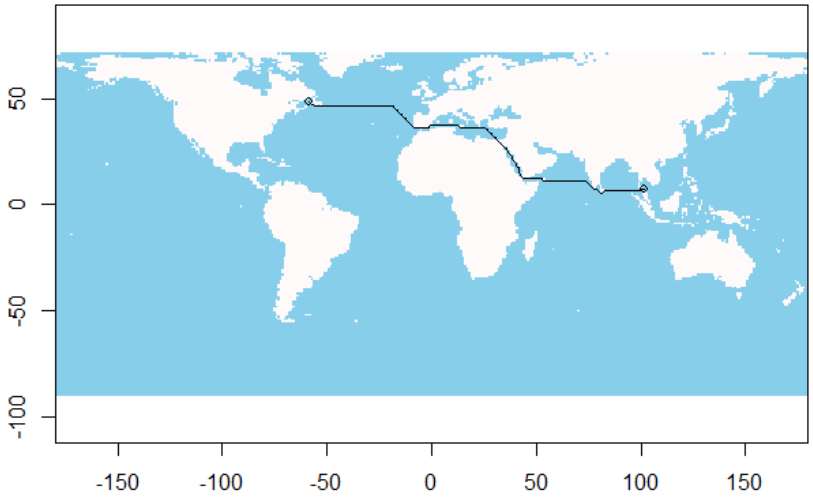


Figure 1.8: Example of shortest path between ports in Canada and Thailand

	(1)	(2)	(3)	(4)	(5)	(6)
centrality$^{\text{cores}}$	-0.03					
	(0.05)					
centrality$^{\text{domestic cores}}$			0.03^{*}		0.04	
			(0.02)		(0.02)	
centrality$^{\text{foreign cores}}$			-0.31^{**}			
			(0.14)			
centrality$^{\text{foreign cores, no FTA}}$					-0.31^{*}	
					(0.16)	
centrality$^{\text{foreign cores, FTA}}$					-0.11	
					(0.12)	
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{core})$		0.33^{**}				
		(0.14)				
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$		-0.05				
		(0.05)				
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$		-0.02				
		(0.07)				
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{core})$				0.31^{***}		0.31^{**}
				(0.11)		(0.15)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$				0.02		0.02
				(0.02)		(0.02)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$				0.05^{**}		0.06^{*}
				(0.02)		(0.03)
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{core})$				-0.18		
				(0.13)		
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$				-0.32^{**}		
				(0.14)		
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$				-0.33^{**}		
				(0.15)		
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{core})$						-0.23
						(0.25)
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$						-0.30
						(0.19)
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{periphery})$						-0.30^{**}
						(0.15)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{core})$						-0.06
						(0.12)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$						-0.11
						(0.11)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{periphery})$						-0.12
						(0.12)
Num. obs.	1460	1460	1460	1460	1392	1392
Num. groups: code	101	101	101	101	97	97
Num. groups: cluster id	3	3	3	3	3	3
Adj. R^2 (proj model)	0.30	0.30	0.31	0.32	0.32	0.33

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. The market potential and centrality variables are entered with the logarithm in the regressions. MP$^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. MP$^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures.

Table 1.7: Regional Development, the Core and Periphery, and Centrality to cores

	(1)	(2)	(3)	(4)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{core})$	-0.33	0.21	14.64	-264.63^{***}
	(0.39)	(0.27)	(18.39)	(36.35)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$	-0.02	0.05^{*}	-0.01	0.00
	(0.03)	(0.03)	(0.02)	(0.06)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$	0.00	0.12^{**}	0.07	-0.09^{**}
	(0.02)	(0.05)	(0.05)	(0.03)
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{core})$	-0.08	0.05	0.03	-2.97^{***}
	(0.36)	(0.45)	(0.27)	(0.59)
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$	-0.53^{**}	0.11	-0.25	-3.24^{***}
	(0.23)	(0.27)	(0.30)	(0.47)
centrality$^{\text{foreign cores, no FTA}} \times \mathbb{1}(\gamma_g = \text{periphery})$	-0.58^{**}	-0.01	-0.15	-3.59^{***}
	(0.23)	(0.20)	(0.43)	(0.60)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{core})$	0.13	-0.20	0.19^{***}	0.34
	(0.13)	(0.20)	(0.06)	(0.25)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$	0.13	-0.24	0.09^{*}	0.29
	(0.14)	(0.15)	(0.05)	(0.20)
centrality$^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{periphery})$	0.11	-0.24	0.03	0.08
	(0.14)	(0.15)	(0.05)	(0.16)
Countries' income sample	high	upper-midde	lower-middle	low
Num. obs.	441	489	346	116
Num. groups: code	29	28	28	12
Num. groups: cluster id	3	3	3	3
Adj. R^2 (full model)	0.84	0.68	0.79	0.97
Adj. R^2 (proj model)	0.25	0.41	0.34	0.45

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. The market potential and centrality variables are entered with the logarithm in the regressions. $\text{MP}^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $\text{MP}^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures.

Table 1.8: Regional Development, the Core and Periphery, and Centrality to cores by countries' income group

	(1)	(2)
$\text{dist}_{ij}^{(\text{haversine})}$	-1.18^{***}	
	(0.02)	
language_{ij}	0.71^{***}	0.66^{***}
	(0.04)	(0.04)
contiguity_{ij}	1.13^{***}	1.12^{***}
	(0.05)	(0.05)
colony_{ij}	1.34^{***}	1.37^{***}
	(0.07)	(0.07)
rta_{ij}	0.40^{***}	0.47^{***}
	(0.04)	(0.03)
currency_{ij}	-0.92^{***}	-0.81^{***}
	(0.11)	(0.11)
$\mathbb{1}_{\text{maritime route}} \times \text{dist}_{io}^{(\text{land, from i to origin port})}$		-0.07^{***}
		(0.01)
$\mathbb{1}_{\text{maritime route}} \times \text{dist}_{od}^{(\text{sea, between ports})}$		-0.96^{***}
		(0.02)
$\mathbb{1}_{\text{maritime route}} \times \text{dist}_{dj}^{(\text{land, from destination port to j})}$		-0.06^{***}
		(0.01)
$(1 - \mathbb{1}_{\text{maritime route}}) \times \text{dist}_{ij}^{(\text{land})}$		-1.00^{***}
		(0.02)
Fixed Effects		
region i	Yes	Yes
region j	Yes	Yes
Num. obs.	$2,378,212$	$2,378,212$
Num. groups: region i	$1,617$	$1,617$
Num. groups: region j	$1,617$	$1,617$
Adj. R^2 (proj model)	0.33	0.33

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. In parenthesis, the robust standard errors are two-way clustered at the regional level.

Table 1.9: Gravity Equation Estimates - regional level

	(1)	(2)	(3)	(4)
$\text{dist}_{c_1c_2}^{(\text{haversine})}$	-1.21^{***}	-1.40^{***}	-1.17^{***}	-1.38^{***}
	(0.09)	(0.07)	(0.08)	(0.06)
$\text{language}_{c_1c_2}$	0.92^{***}	0.82^{***}	0.87^{***}	0.72^{***}
	(0.16)	(0.11)	(0.21)	(0.10)
$\text{contiguity}_{c_1c_2}$	0.86^{***}	0.72^{***}	0.81^{***}	0.70^{***}
	(0.23)	(0.21)	(0.14)	(0.19)
$\text{colony}_{c_1c_2}$	1.02^{***}	0.65^{***}	1.18^{***}	0.66^{***}
	(0.29)	(0.15)	(0.29)	(0.14)
$\text{rta}_{c_1c_2t}$	0.32^{**}	0.51^{***}	0.32^{**}	0.44^{***}
	(0.14)	(0.11)	(0.12)	(0.10)
$\text{currency}_{c_1c_2t}$	-0.83^{**}	-0.04	-0.94^{***}	-0.11
	(0.38)	(0.35)	(0.31)	(0.28)
Fixed Effects				
Country c_1	Yes	Yes	No	No
Country c_2	Yes	Yes	No	No
Country c_1 $\times$ year t	No	No	Yes	Yes
Country c_2 $\times$ year t	No	No	Yes	Yes
Sample				
Year	2005	2005	[1996; 2010]	[1996; 2010]
Countries	BACI $\cap$ Gennaioli	All in BACI	BACI $\cap$ Gennaioli	All in BACI
Num. obs.	7,057	18,618	99,928	254,889
Num. groups: $c_1(\times t)$	77	146	1,155	2,190
Num. groups: $c_2(\times t)$	103	208	1,545	3,116
Adj. R^2 (proj model)	0.30	0.26	0.30	0.25

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. In parenthesis, the robust standard errors are two-way clustered at the national level. Note that the four first variables in the estimation are time invariant in the period [1996; 2010], while the two latter are time varying for some pairs of countries.

Table 1.10: Gravity Equation Estimates - national level

	HM 1	HM 2
$\text{dist}_{ij}^{(\text{haversine})}$	-0.93	-1.10
language_{ij}	0.54	0.39
contiguity_{ij}	0.53	0.66
colony_{ij}	0.92	0.75
rta_{ij}	0.59	0.75
currency_{ij}	0.79	0.86
home_{ij}	1.96	1.90

Average trade elasticity to resistance and facilitator variables found in the literature, gathered in the meta-analysis by Head and Mayer (2014). First column displays the average coefficients in the whole sample of gravity models in their analysis, while the second only includes coefficients estimated from structural gravity models.

Table 1.11: Gravity Equation Estimates - Head and Mayer (2014)

	$\beta^{(1)} = \hat{\beta}_{C1}$	$\beta^{(2)} = \hat{\beta}_{C1}$	$\beta_{C1} = \hat{\beta}_{HM1}$	$\beta_{C1} = \hat{\beta}_{HM2}$	$\beta^{(1)} = \hat{\beta}_{HM1}$	$\beta^{(1)} = \hat{\beta}_{HM2}$
$\text{dist}_{ij}^{(\text{haversine})}$	0.23	-	0.00	0.24	0.00	0.00
language_{ij}	0.00	0.00	0.01	0.00	0.00	0.00
contiguity_{ij}	0.00	0.00	0.15	0.36	0.00	0.00
colony_{ij}	0.00	0.00	0.83	0.44	0.00	0.00
rta_{ij}	0.00	0.00	0.03	0.62	0.00	0.32
currency_{ij}	0.28	0.95	0.00	0.00	0.00	0.00

The table presents p-values of the tests related to the null hypothesis displayed in the top of each column. A p-value of $p < 0.05$ rejects the null hypothesis at the 95% confidence level. If $p \geq 0.05$, we can consider that the null hypothesis is not rejected.

Table 1.12: Statistical tests - p-values

	All countries					*High income countries*				
	mean	**min**	**max**	**sd**	**n**	**mean**	**min**	**max**	**sd**	**n**
log(GDP pc)	8.77	1.76	11.87	1.32	1502	10.02	8.9	11.24	0.45	289
temperature	15	-12.73	29.15	8.45	1501	10	-12.2	27.27	4.94	289
inv. dist. port	0.04	0	1	0.18	1502	0.07	0	1	0.23	289
log(oil production pc)	0.1	0	4.16	0.4	1502	0.05	0	3.13	0.27	289
log(density)	4.17	-2.93	12.01	1.81	1502	4.79	-2.93	10.23	1.55	289
years education	7.13	0.22	13.21	3.16	1465	10.09	5.3	12.94	1.71	289
log($MP^{(h)}$)	21.21	19.07	25.49	1.32	1502	22.74	20.6	25.49	0.93	289
log($MP^{(s)}$)	22.23	20.06	25.92	1.22	1502	23.62	21.81	25.92	0.89	289
log($NLMP^{(h)}$)	21.18	19.06	25.46	1.31	1502	22.69	20.55	25.46	0.94	289
log($NLMP^{(s)}$)	22.21	20.04	25.9	1.21	1502	23.59	21.74	25.9	0.89	289
log($FMP^{(h)}$)	20.53	19.06	24.33	1.05	1502	21.61	19.32	24.33	1.29	289
log($FMP^{(s)}$)	21.62	20.03	25.2	0.98	1502	22.59	20.19	25.2	1.29	289
	Upper middle income countries					*Lower middle income countries*				
	mean	**min**	**max**	**sd**	**n**	**mean**	**min**	**max**	**sd**	**n**
log(GDP pc)	8.9	7.35	11.12	0.55	485	7.84	06.09	9.75	0.57	393
temperature	14.69	-12.73	27.99	8.66	485	19	-5.79	28.26	8.18	393
inv. dist. port	0.05	0	1	0.2	485	0.02	0	1	0.11	393
log(oil production pc)	0.23	0	3.48	0.59	485	0.02	0	1.98	0.12	393
log(density)	3.74	-2.67	10.43	1.82	485	4.24	-1.91	9.56	1.78	393
years education	6.89	2.37	11.45	1.83	485	5.05	0.44	12.5	2.42	393
log($MP^{(h)}$)	20.95	19.54	25.04	0.76	485	20.31	19.24	23.18	0.53	393
log($MP^{(s)}$)	21.98	20.75	25.45	0.69	485	21.4	20.31	23.68	0.47	393
log($NLMP^{(h)}$)	20.91	19.52	24.91	0.75	485	20.3	19.24	23.18	0.53	393
log($NLMP^{(s)}$)	21.96	20.75	25.34	0.68	485	21.39	20.31	23.68	0.47	393
log($FMP^{(h)}$)	20.3	19.21	22.03	0.66	485	20.13	19.19	21.86	0.46	393
log($FMP^{(s)}$)	21.42	20.59	23.05	0.6	485	21.26	20.17	22.78	0.44	393
	Low income countries									
	mean	**min**	**max**	**sd**	**n**					
log(GDP pc)	6.38	1.76	8.56	1.37	134					
temperature	22.8	-1.58	29.15	6.19	134					
inv. dist. port	0.03	0	1	0.17	134					
log(oil production pc)	0	0	0.08	0.01	134					
log(density)	4.29	-0.55	12.01	1.73	134					
years education	2.98	0.22	11.7	2.5	134					
log($MP^{(h)}$)	19.7	19.07	20.58	0.25	134					
log($MP^{(s)}$)	20.85	20.06	21.34	0.23	134					
log($NLMP^{(h)}$)	19.7	19.06	20.58	0.25	134					
log($NLMP^{(s)}$)	20.85	20.04	21.34	0.23	134					
log($FMP^{(h)}$)	19.67	19.06	20.22	0.24	134					
log($FMP^{(s)}$)	20.83	20.03	21.27	0.22	134					

Country income groups are defined following the World Bank classification: *High income* (21 countries in the sample), *Upper middle income* (28 countries), *Lower middle income* (31 countries) and *Low income* (14 countries).

Table 1.13: Descriptive statistics - 2005

	core					*semi-periphery*				
	mean	**min**	**max**	**sd**	**n**	**mean**	**min**	**max**	**sd**	**n**
log(GDP pc)	9.11	2.63	11.42	1.31	142	8.76	1.98	11.24	1.37	379
temperature	16.69	-4.66	29.1	7.85	142	15.69	-9.1	29.15	7.95	379
inv. dist. port	0.08	0	1	0.24	142	0.06	0	1	0.22	379
log(oil per cap.)	0.09	0	3.48	0.45	142	0.05	0	2.42	0.24	379
log(density)	6.17	1.42	11.1	1.86	142	4.77	-0.23	12.01	1.46	379
years education	7.84	0.25	13.21	2.98	142	7.1	0.27	12.85	3.29	379
log($MP^{(h)}$)	21.35	19.09	25.04	1.47	142	21.25	19.07	25.49	1.4	379
log($MP^{(s)}$)	22.29	20.13	25.45	1.31	142	22.26	20.06	25.92	1.28	379
log($NLMP^{(h)}$)	21.17	19.09	24.91	1.42	142	21.21	19.06	25.46	1.39	379
log($NLMP^{(s)}$)	22.19	20.13	25.34	1.28	142	22.24	20.04	25.9	1.28	379
log($FMP^{(h)}$)	20.57	19.08	24.27	1.14	142	20.6	19.06	24.33	1.12	379
log($FMP^{(s)}$)	21.69	20.13	25.17	1.05	142	21.69	20.03	25.2	1.05	379
	periphery									
	mean	**min**	**max**	**sd**	**n**					
log(GDP pc)	8.71	1.76	11.87	1.3	977					
temperature	14.48	-12.73	28.8	8.69	977					
inv. dist. port	0.03	0	1	0.14	977					
log(oil per cap.)	0.12	0	4.16	0.44	977					
log(density)	3.65	-2.93	9.17	1.64	977					
years education	07.03	0.22	12.83	3.12	977					
log($MP^{(h)}$)	21.18	19.13	24.88	1.27	977					
log($MP^{(s)}$)	22.21	20.09	25.54	1.19	977					
log($NLMP^{(h)}$)	21.16	19.12	24.62	1.26	977					
log($NLMP^{(s)}$)	22.2	20.09	25.44	1.18	977					
log($FMP^{(h)}$)	20.49	19.12	24.2	1	977					
log($FMP^{(s)}$)	21.58	20.08	25.11	0.95	977					

Table 1.14: Descriptive statistics by cluster characteristic

	(1)	(2)	(3)	(4)
	ln education		ln density	
$\mathbb{1}_{\text{semi-periphery}}$	-0.18^{***}	-0.04^{***}	-1.42^{***}	-1.13^{***}
	(0.03)	(0.01)	(0.16)	(0.28)
$\mathbb{1}_{\text{periphery}}$	-0.27^{***}	-0.07^{***}	-2.42^{***}	-1.99^{***}
	(0.04)	(0.02)	(0.16)	(0.29)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{upper-middle income group}}$		-0.09^{***}		-0.51
		(0.03)		(0.41)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{upper-middle income country}}$		-0.16^{***}		-0.88^{**}
		(0.03)		(0.39)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{lower-middle income country}}$		-0.17^{***}		-0.14
		(0.06)		(0.40)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{lower-middle income country}}$		-0.28^{***}		-0.24
		(0.09)		(0.40)
$\mathbb{1}_{\text{semi-periphery}} \times \mathbb{1}_{\text{low income country}}$		-0.55^{***}		-0.89
		(0.17)		(0.68)
$\mathbb{1}_{\text{periphery}} \times \mathbb{1}_{\text{low income country}}$		-0.67^{***}		-0.82
		(0.19)		(0.55)
Num. obs.	1479	1479	1516	1516
Num. groups: code	102	102	104	104
Adj. R^2 (full model)	0.90	0.91	0.54	0.55
Adj. R^2 (proj model)	0.14	0.20	0.27	0.28

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. The dependant variable is the GDP per capita (in log). Robust standard errors adjusted for clustering on each country are in parentheses. Estimations include country fixed-effects.

Table 1.15: Core and Periphery Divide, Education and Density

	(1)	(2)	(3)	(4)	(5)	(6)
	ln $MP^{(h)}$	ln $MP^{(s)}$	ln $NLMP^{(h)}$	ln $NLMP^{(s)}$	ln $FMP^{(h)}$	ln $FMP^{(s)}$
$\mathbb{1}_{\text{semi-periphery}}$	-0.20^{***}	-0.13^{***}	-0.05	-0.05	-0.00	-0.02^{*}
	(0.06)	(0.05)	(0.06)	(0.05)	(0.01)	(0.01)
$\mathbb{1}_{\text{periphery}}$	-0.36^{***}	-0.25^{***}	-0.18^{***}	-0.15^{***}	-0.02^{**}	-0.05^{***}
	(0.07)	(0.06)	(0.07)	(0.05)	(0.01)	(0.01)
Num. obs.	1498	1498	1498	1498	1498	1498
Num. groups: code	103	103	103	103	103	103
Adj. R^2 (full model)	0.91	0.93	0.91	0.93	0.98	0.98
Adj. R^2 (proj model)	0.07	0.05	0.03	0.03	0.00	0.01

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. The dependant variable is the GDP per capita (in log). Robust standard errors adjusted for clustering on each country are in parentheses. Estimations include country fixed-effects.

Table 1.16: Core and Periphery Divide, Market Potential

	(1)	(2)	(3)	(4)	(5)	(6)
market potential	0.22^{***}	0.22^{***}	0.14^{***}	0.15^{***}	0.13	0.12
	(0.05)	(0.06)	(0.04)	(0.05)	(0.11)	(0.13)
Num. obs.	1,502	1,502	1,502	1,502	1,502	1,502
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Num. groups: code	105	105	105	105	105	105
Adj. R^2 (proj model)	0.06	0.04	0.02	0.01	0.00	0.00
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. The table displays the estimated coefficients from equation 1.6. Robust standard errors adjusted for clustering on each country are in parentheses. $MP^{(h)}$ considers physical distance between regions measured as the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDP_i\tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j\neq i} GDP_j\tau_{ij}$ with i and j both in country c.

Table 1.17: Regional Development and Market potential - Univariate regressions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
market potential	0.16^{**}	0.16^{**}	0.11^{**}	0.11^{**}	0.03	-0.02
	(0.06)	(0.07)	(0.05)	(0.06)	(0.12)	(0.10)
education +65 years old	0.22^{***}	0.22^{***}	0.22^{***}	0.22^{***}	0.22^{***}	0.22^{***}
	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)
Num. obs.	607	607	607	607	607	607
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Num. groups: code	39	39	39	39	39	39
Adj. R^2 (proj model)	0.38	0.37	0.37	0.37	0.36	0.36
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses.

Table 1.18: Regional development and Market Potential - Education of old

	(1)	(2)	(3)	(4)	(5)	(6)
market potential	0.12^{***}	0.08^{**}	0.05^{**}	0.04	0.05	-0.02
	(0.04)	(0.03)	(0.02)	(0.03)	(0.07)	(0.09)
centrality	0.05^{***}	0.05^{***}	0.06^{***}	0.06^{***}	0.06^{***}	0.06^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Num. obs.	1464	1464	1464	1464	1464	1464
Num. groups: code	103	103	103	103	103	103
Adj. R^2 (full model)	0.95	0.95	0.95	0.95	0.95	0.95
Adj. R^2 (proj model)	0.44	0.43	0.43	0.43	0.43	0.42
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Market potential and centrality indexes are introduced in the logarithm form. In each regression, I control for the temperature, the inverse distance to the closest port, oil production per capita, density and average level of education. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDP_i\tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j\neq i} GDP_j\tau_{ij}$ with i and j both in country c.

Table 1.19: Market Potential and Centrality (2005)

	(1)	(2)	(3)	(4)	(5)	(6)
market potential	0.17^{***}	0.17^{***}	0.12^{***}	0.13^{***}	0.11^{*}	0.05
	(0.04)	(0.04)	(0.03)	(0.04)	(0.06)	(0.06)
market potential $\times \mathbb{1}_{\text{semi-periphery}}$	-0.04^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
market potential $\times \mathbb{1}_{\text{periphery}}$	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
centrality	-0.00	-0.00	0.00	0.00	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
centrality $\times \mathbb{1}_{\text{semi-periphery}}$	-0.17^{***}	-0.16^{***}	-0.15^{***}	-0.15^{***}	-0.11^{***}	-0.11^{***}
	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
centrality $\times \mathbb{1}_{\text{periphery}}$	-0.12^{**}	-0.13^{**}	-0.10^{*}	-0.11^{**}	-0.07	-0.07
	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)
Num. obs.	1460	1460	1460	1460	1460	1460
Num. groups: code	101	101	101	101	101	101
Adj. R^2 (full model)	0.95	0.95	0.95	0.95	0.95	0.95
Adj. R^2 (proj model)	0.48	0.48	0.47	0.4708	0.46	0.46
Regressor	$MP^{(h)}$	$MP^{(s)}$	$NLMP^{(h)}$	$NLMP^{(s)}$	$FMP^{(h)}$	$FMP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Market potential and centrality indexes are introduced in the logarithm form. In each regression, I control for the temperature, the inverse distance to the closest port, oil production per capita, density and average level of education. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $NLMP^{(h)}$ and $NLMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the local market, i.e. $GDP_i\tau_{ii}$. $FMP^{(h)}$ and $FMP^{(s)}$ are proxies for $MP^{(h)}$ and $MP^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j\neq i} GDP_j\tau_{ij}$ with i and j both in country c.

Table 1.20: Market Potential and Centrality - Core and Periphery (2005)

	(1)	(2)	(3)	(4)	(5)	(6)
1995 sample						
market potential	0.10^{**}	0.08^{*}	0.03	0.02	0.13^{*}	0.08
	(0.05)	(0.05)	(0.04)	(0.04)	(0.07)	(0.09)
2000 sample						
market potential	0.10^{**}	0.09^{*}	0.04	0.04	0.22^{***}	0.18^{*}
	(0.05)	(0.05)	(0.04)	(0.05)	(0.07)	(0.09)
2005 sample						
market potential	0.11^{***}	0.09^{**}	0.06^{**}	0.05^{*}	0.08	-0.00
	(0.03)	(0.04)	(0.03)	(0.03)	(0.08)	(0.10)
2010 sample						
market potential	0.01	-0.07^{*}	-0.06	-0.11^{**}	-0.05	-0.23^{**}
	(0.03)	(0.04)	(0.03)	(0.03)	(0.08)	(0.10)
1995-2005 sample						
market potential	0.10^{***}	0.08^{***}	0.04^{**}	0.03	0.13^{***}	0.06
	(0.02)	(0.03)	(0.02)	(0.02)	(0.05)	(0.06)
1995-2010 sample						
market potential	0.08^{**}	0.05^{*}	0.02	0.00	0.09^{**}	-0.01
	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)	(0.05)
Regressor	$\text{MP}^{(h)}$	$\text{MP}^{(s)}$	$\text{NLMP}^{(h)}$	$\text{NLMP}^{(s)}$	$\text{FMP}^{(h)}$	$\text{FMP}^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country-time group are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. For the cross-sectional estimations of each year, country fixed effects are included. For the panel estimations, country-year fixed effects are included. $\text{MP}^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $\text{MP}^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $\text{NLMP}^{(h)}$ and $\text{NLMP}^{(s)}$ are proxies for $\text{MP}^{(h)}$ and $\text{MP}^{(s)}$ respectively, excluding the local market, i.e. $\text{GDP}_i\tau_{ii}$. $\text{FMP}^{(h)}$ and $\text{FMP}^{(s)}$ are proxies for $\text{MP}^{(h)}$ and $\text{MP}^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j\neq i}\text{GDP}_j\tau_{ij}$ with i and j both in country c.

Countries in the 1995 sample: ARG, SVN, NOR, PAK, PAN, PER, POL, PRT, SLV, SWE, THA, TUR, TZA, URY, AUS, AUT, BEL, BGR, BOL, BRA, CAN, CHE, CHL, CHN, COL, DEU, ECU, ESP, FIN, FRA, GRC, GTM, HND, IDN, IND, IRL, ITA, JOR, JPN, KAZ, LKA, LSO, LTU, LVA, MEX, MNG, MYS, NLD. The sample gathers 738 regions in 48 countries.

Countries in the 2000 sample: ARG, SVN, NOR, PAN, PER, POL, PRT, PRY, SLV, SRB, SWE, THA, TUR, TZA, URY, ALB, ARE, AUS, AUT, BEL, BEN, BGR, BOL, BRA, CANn CHE, CHL, CHN, COL, DEU, ECU, ESP, EST, FIN, FRA, GRC, HND, HRV, HUN, IND, IRL, ITA, JOR, JPN, KAZ, LKA, LSO, LTU, LVA, MEX, MKD, MNG, MOZ, MYS, NLD. The sample gathers 802 regions in 55 countries.

Countries in the 2005 sample: ARG, SVN, NOR, NPL, PAK, PAN, PER, POL, PRT, PRY, RUS, SLV, SVK, SWE, THA, TUR, TZA, UKR, URY, USA, UZB, VNM, ARE, AUS, AUT, BEL, BEN, BGR, BIH, BOL, BRA, CAN, CHE, CHL, CHN, COL, CZE, DEU, ECU, ESP, EST, FIN, FRA, GBR, GRC, GTM, HND, HRV, HUN, IDN, IND, IRL, IRN, ITA, JOR, JPN, KAZ, KGZ, LKA, LSO, LTU, LVA, MEX, MKD, MNG, MOZ, MYS, NGA, NLD. The sample gathers 1129 regions in 69 countries.

Countries in the sample: SVN, NOR, PAN, PRT, RUS, SVK, SWE, USA, VNM, BGR, BRA, CHE, CHN, CZE, FIN, IDN, JPN, KOR, MEX, NLD. The sample gathers 421 regions in 20 countries.

Table 1.21: Market potential elasticity coefficients - panel

	(1)	(2)	(3)	(4)	(5)	(6)
1995-2005 sample						
market potential	0.09^{**}	0.08^{**}	0.05	0.05	0.11^{*}	0.06
	(0.04)	(0.04)	(0.03)	(0.03)	(0.06)	(0.07)
market potential $\times \mathbb{1}_{\text{semi-periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times \mathbb{1}_{\text{periphery}}$	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.03^{***}	-0.03^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1995-2010 sample						
market potential	0.08^{**}	0.05	0.03	0.01	0.08	0.00
	(0.04)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)
market potential $\times \mathbb{1}_{\text{semi-periphery}}$	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}	-0.01^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
market potential $\times \mathbb{1}_{\text{periphery}}$	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.03^{***}	-0.03^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Regressor	$\text{MP}^{(h)}$	$\text{MP}^{(s)}$	$\text{NLMP}^{(h)}$	$\text{NLMP}^{(s)}$	$\text{FMP}^{(h)}$	$\text{FMP}^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country-time group are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. For the cross-sectional estimations of each year, country fixed effects are included. For the panel estimations, country-year fixed effects are included. $\text{MP}^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $\text{MP}^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures. $\text{NLMP}^{(h)}$ and $\text{NLMP}^{(s)}$ are proxies for $\text{MP}^{(h)}$ and $\text{MP}^{(s)}$ respectively, excluding the local market, i.e. $\text{GDP}_i \tau_{ii}$. $\text{FMP}^{(h)}$ and $\text{FMP}^{(s)}$ are proxies for $\text{MP}^{(h)}$ and $\text{MP}^{(s)}$ respectively, excluding the domestic markets, i.e. $\sum_{j \neq i} \text{GDP}_j \tau_{ij}$ with i and j both in country c.

Table 1.22: Market potential elasticity coefficients - panel - core-periphery

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
market potential	0.14^{***}	0.12^{***}	0.15^{***}	0.12^{***}	0.15^{***}	0.13^{**}	0.16^{***}	0.13^{**}
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)
centrality$^{\text{cores}}$	-0.11	-0.09						
	(0.07)	(0.06)						
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{core})$			-0.37	-0.12				
			(0.26)	(0.22)				
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$			-0.13^{**}	-0.10^{*}				
			(0.06)	(0.06)				
centrality$^{\text{cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$			-0.10	-0.08				
			(0.08)	(0.08)				
centrality$^{\text{domestic cores}}$					-0.00	0.01		
					(0.02)	(0.02)		
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{core})$							-0.37	-0.12
							(0.28)	(0.24)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$							-0.02	-0.01
							(0.02)	(0.02)
centrality$^{\text{domestic cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$							0.01	0.03
							(0.02)	(0.02)
centrality$^{\text{foreign cores}}$					-0.42^{**}	-0.39^{**}		
					(0.18)	(0.18)		
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{core})$							-0.32^{*}	-0.27
							(0.18)	(0.17)
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$							-0.43^{**}	-0.40^{**}
							(0.17)	(0.17)
centrality$^{\text{foreign cores}} \times \mathbb{1}(\gamma_g = \text{periphery})$							-0.44^{**}	-0.41^{**}
							(0.19)	(0.19)
Num. obs.	1460	1460	1460	1460	1460	1460	1460	1460
Country and Cluster FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num. groups: code	101	101	101	101	101	101	101	101
Adj. R^2 (proj model)	0.32	0.31	0.32	0.31	0.33	0.32	0.34	0.33
Regressor	$MP^{(h)}$	$MP^{(s)}$	$MP^{(h)}$	$MP^{(s)}$	$MP^{(h)}$	$MP^{(s)}$	$MP^{(h)}$	$MP^{(s)}$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors adjusted for clustering on each country are in parentheses. The following covariates are included: temperature, inverse distance to the closest port, oil production per capita and average educational level. The market potential and centrality variables are entered with the logarithm in the regressions. $MP^{(h)}$ considers physical distance between regions measured by the haversine distance, cultural proximity measures depending on common language, national common border, colonial ties, and trade facilities implied by a common currency and regional trade agreements. $MP^{(s)}$ considers the shortest path by land and sea between regions, passing through their closest ports if needed, in addition to the cultural proximity and trade facilities measures.

Table 1.23: Regional Development, the Core and Periphery, and Centrality to cores (2)

	(1)	(2)	(3)	(4)
$\text{centrality}^{\text{foreign cores, FTA}}$	0.03	0.04		
	(0.06)	(0.06)		
$\text{centrality}^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{semi-periphery})$		−0.00	0.03	0.01
		(0.02)	(0.06)	(0.08)
$\text{centrality}^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{periphery})$		−0.00	0.03	0.01
		(0.02)	(0.06)	(0.08)
$\text{centrality}^{\text{foreign cores, FTA}} \times \mathbb{1}(\gamma_g = \text{core})$			0.04	0.02
			(0.06)	(0.08)
Num. obs.	2291	2291	2291	1563
region FE	890	890	890	674
cluster group FE		3	3	3
country × year FE	143	143	143	103
Covariates	No	No	No	Yes
Adj. R^2 (full model)	0.99	0.99	0.99	0.99
Adj. R^2 (proj model)	−0.00	−0.00	−0.00	0.22

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors are clustered at the region level.

Table 1.24: Regional Development and Trade Agreement - Panel

Chapter 2

High-Speed Railways and the Geography of Inventors' Collaboration: Evidence from France (1980-2010)

JOINT WITH FERNANDO STIPANICIC

Abstract

This study explores the impact of high-speed railways (HSR) on inventor collaboration over long distances, which plays as a catalyst of face-to-face interactions and knowledge exchange. We use a novel region-to-region travel time dataset using HSR implementation in France. Employing a gravity model with three-way fixed effects, we assess the causal relationship between travel time reduction and cross-regional co-patenting trends between NUTS3 region-pairs, addressing endogeneity concerns. Results show a robust positive effect of reduced travel time on collaboration, most of all as comparison of intra-regional and long-distance collaborations. Core regions significantly benefit, but advantages extend beyond directly-HSR-connected regions. Moreover, reduced travel time is related to collaborative patents exhibiting higher novelty and wider scope within the realm of technology fields. Finally, we find that the reduction in travel time has fostered connections among inventor leaders and inventors with superior productivity compared to their collaborators' pool.

Keywords: High-Speed Railways, Collaboration, Patents, Gravity
JEL Classification: O18, O34, O36, R40

2.1 Introduction

French inventions applied at the European Patent Office reveal a substantial rise in collaborative patenting over time. Collaborative patents have witnessed a significant increase, from representing 36% of all patents in 1980 to reaching 62% by 2010. This trend is primarily driven by the growth of inter-regional co-patents, which have shown a remarkable surge, outpacing the increase in intra-regional co-patents. Specifically, inter-regional co-patents experienced a staggering growth rate of 104%, while intra-regional co-patents grew at 53% over the same period. As a result, the gap between intra-regional and inter-regional co-patents has narrowed, making long-distance collaborations increasingly more prevalent in collaborative inventions.

Another significant aspect is the geographical distance between collaborators. In 1980, the average distance between collaborators within patent team was approximately 60 kilometers. However, as time went by, this average distance has been extended by 30 additional kilometers by 2010. For inter-regional co-patents specifically, the average distance between collaborators has consistently risen from 110 kilometers to 140 kilometers over the period.

During this time period, France experienced a significant expansion of its high-speed rail network, with the inauguration of the first line connecting Paris and Lyon in 1981. This development drastically reduced the travel time for this 400-kilometer journey from 3 hours and 40 minutes to 1 hour and 40 minutes, making it feasible for individuals to undertake the trip even twice a day, if needed. The mentioned observed trends of rising collaboration and the broader geographic scope of collaborative inventions provide an opportunity to investigate the influence of transportation enhancements on collaborative interactions.

This paper investigates the impact of transportation infrastructure improvements on long-distance interactions. We specifically focus on the roll-out of high-speed railways (HSR) as a quasi-natural experiment, due to their transformative effect on the relationship between space and time. Unlike transportation systems primarily geared towards goods, high-speed railways cater to transporting people, making them particularly influential in promoting face-to-face interactions over economic exchange. While directly observing face-to-face interactions can be challenging, we can identify and investigate their economic significance through the lens of innovation collaboration, notably in the context of patent collaboration.

Our hypothesis revolves around the idea that increased distance intensifies the challenges and costs associated with discovering new collaborators and managing ongoing collaborations. Since high-speed railways reduce travel time between locations, it can facilitate long-distance collaboration by providing additional time for in-person collaborative efforts on research projects. We posit that their introduction will lead to heightened degree of collaboration among inventors in connected regions.

The present paper encompasses four main analysis. The first delves into the impact of travel time reduction on the quantity of co-patents between regions. The second part scrutinizes heterogeneous effects based on HSR connectivity, distance, and the economic and innovative importance of regions. The third part explores effects on collaborative innovation at both the intensive and extensive margins of collaboration, along with the impact on the quality of collaborative innovation. The final part investigates the mechanisms behind HSR connectivity, particularly focusing on what type of inventors the HSR connects. To do so, we classify inventors according to their productivity, drawing from Akcigit et al. (2018) and Catalini et al. (2020).

The rationale for focusing on France in this study arises from the four-decade presence

of high-speed railways in the country. This extended period provides an ideal opportunity to observe the long-term effects of enhanced connectivity on innovation output. Furthermore, France holds a prominent position as one of the global leaders in innovation, currently ranking 6th in the world and among the top in Europe in terms of R&D expenditure, as highlighted by the Global Innovation Index 2022 (GII). Recognizing the pivotal role of innovation in driving economic growth and development, France has proactively pursued innovation policies aimed at nurturing its research ecosystem.

In particular, innovation policies in France have underscored the significance of knowledge transfer and collaboration between the scientific and industrial sectors. Approximately 15% of these policy initiatives are designed to facilitate the exchange of knowledge between science and industry, with an additional 50% dedicated to promoting collaborative research endeavors. The prevalence of policy initiatives on collaborative research in France is underscored by its association with physical cluster creation, as evident in the data by Science, Technology and Innovation Policy (STIP).[1] This emphasis on collaboration and clustering reflects a strategic approach to gather individuals and firms of similar activity to work together and share resources and knowledge in order to leverage the collective expertise of researchers, industries, and institutions in driving innovation and achieving global competitiveness (Kerr and Robert-Nicoud, 2020; Moretti, 2021).

However, while clusters offer numerous benefits in terms of collaboration and knowledge exchange, they can also lead to a lock-in effect, as highlighted by Boschma (2005). The lock-in effect suggests that after a certain period of close collaboration, it becomes increasingly challenging for individuals and firms to acquire new knowledge and to make significant progress on innovative projects. This phenomenon can limit the diversity of ideas and perspectives, potentially hindering further innovation and growth.

Catalini et al. (2020) have presented compelling evidence that better transportation connectivity facilitates greater interaction among researchers from diverse geographical locations, which results in the opportunity for inventors to collaborate with distant ones with higher productivity than inventors in their local pools. Those collaborations are more likely to yield fruitful outcomes, enhancing the overall quality of research and fostering innovation. As evidenced by Akcigit et al. (2018), improved access to distant regions positively influence inventors' productivity, and ultimately drives economic growth through higher-quality patents.

Physical proximity has long been recognized as a key factor in fostering knowledge diffusion and innovation. For instance, Jaffe et al. (1993) found that in the United States, citations to domestic patents are more likely to be from domestic sources. Inventor collaboration, which is essential for innovation, typically occurs among individuals who are geographically close. Catalini (2018) provides evidence for the significance of colocation in influencing the rate, direction, and quality of scientific collaboration.[2] Numerous studies, including Guellec and van Pottelsberghe de la Potterie (2001), Hoekman et al. (2009), Picci (2010) and Bergé (2015), highlight the negative relationship between collaboration and distance, and emphasize that a significant share of collaborations occurs within national borders in Europe.

Jaffe et al. (1993) report evidence indicating that the localization effect gradually diminishes over time. Developments in transportation infrastructure and communication technolo-

[1] EC-OECD (2023), STIP Compass: International Database on Science, Technology and Innovation Policy (STIP), edition August 4, 2023, https://stip.oecd.org.

[2] The study capitalizes on the constraints placed on lab locations within the Jussieu campus of Paris, driven by the need for asbestos removal from its buildings, offering a unique opportunity to examine these effects.

gies may have had profound influence on reducing interaction costs between geographically distant parties. Indeed, several studies have delved into the impact of transportation advancements on innovation. For instance, Agrawal et al. (2017) investigated the influence of a region's highway network on innovation, or Perlman et al. (2016) who focuses on the effect of railroad networks density.

Our study belongs to the strand of literature that employs transportation network enhancements as quasi-natural experiments on innovative aspects. Among them, we find Tsiachtsiras (2022) and Andersson et al. (2023), who investigated the effects of railroad expansion in France and Sweden, respectively, using historical data dating back to the 19th century, or Pauly and Stipanicic (2022), who estimated the impact of travel time reduction following the introduction of jet transportation in the United States. Over shorter distances, Bernard et al. (2020) assesses the impact of bridges construction between Japan's islands on collaboration, while Koh et al. (2022) examines the case of subway expansion in Beijing.

Our contribution to the literature standed out that we investigate the ability of high-speed railways to connect inventors. However, the principle of multiple discovery did not spare our work, as evidenced by the quantity of very recent papers examining the same topic (Hanley et al., 2022; Li et al., 2022; Yao and Li, 2022; Kang et al., 2023), exclusively for the case of China. Such like them, we study collaboration by using patent data, but unlike them, we identify co-patents between inventors rather than firms. This choice is informed by the availability and appropriateness of European patent data, which has been meticulously cleaned by Morrison et al. (2017) for reporting on individual-level collaborations over time. Importantly, the referenced studies all center on the impact of HSR systems in China, which is particularly relevant due to the rapid expansion of the Chinese HSR network, and detailed travel time data readily available.[3]

One significant contribution of this paper lies in addressing a data gap specific to France. Travel time data is currently accessible for only a very limited number of cities, provided by the Société National des Chemins de Fer, the French public railway company. To fill this gap, we have computed a novel dataset, which includes estimations of station-to-station travel times for every year, starting from the year before the introduction of the first high-speed rail line up to the present day. This unique dataset allows us to capture the changes in travel time that have occurred following the implementation of each high-speed railway line.

We capitalize on the variation in travel time resulting from the expansion and extension of the HSR network in order to evaluate collaboration patterns within region-pairs, using a gravity model with three-way fixed effects. The econometric model controls for a rich set of pair fixed effects and region-time fixed effects, providing a robust evaluation of the relationship. While many studies have relied on difference-in-differences models to examine the impact of direct connections on treated pairs (Hanley et al., 2022; Li et al., 2022; Yao and Li, 2022), our approach differs. Both directly and indirectly connected pairs have experienced a decrease in travel time due to HSR network developments. We leverage this information to estimate improved access across all pairs. Additionally, certain pairs, such as Paris and Lyon, have witnessed reductions in travel time in different years (in 1981, 1983, and 2001), which makes difficult the choose of one date of treatment.

Potential concerns about reverse causality arise if high-speed railways were constructed between cities already collaborating and experiencing co-patent growth, implying a strategic

[3]Other papers have investigated the effect of HSR on innovation, including Dong et al. (2018), Gao and Zheng (2020), and Tsiachtsiras et al. (2022).

motive to amplify their collaborations. However, our approach mitigates these concerns by adopting a inconsequential place approach. We exclude pairs of regions with a direct connection to the HSR network, avoiding issues related to government decisions based on past or predicted collaborations - Kang et al. (2023) has a similar approach. Indeed, those pairs have experienced reduction in travel time by virtue of relative proximity to an HSR station. Especially, we remove pairs of regions with a direct connection to the HSR network, i.e., those with an HSR station.

We address concerns related to omitted variable bias more effectively than most papers closely related to our study. Building on the work of Bergé (2015), we compute a measure called *bridges* in network analysis. This measure quantifies the common collaborators between every pair of inventors, excluding their own collaboration, and aggregates it at the region-pair level. Bridges are individuals who connect two others within the network, facilitating the flow of information and resources, considering that two inventors might have more chance to collaborate in the future if they have a common collaborator. We find that this measure is crucial to consider as it correlates with travel time. As travel time decreases, the interconnectedness of inventors' collaboration network increases. This measure significantly alleviates omitted variable bias, avoiding an upward bias in the estimate (a downward bias in the coefficient's magnitude), thereby improving the precision of the estimator associated to travel time.

We additionally control for regional technological proximity as usually done in the literature, computing the measure following Jaffe (1986). Furthermore, we account for the impact of evolving internet connectivity, which could potentially compete with the need for face-to-face interactions. We leverage data on ADSL geographical coverage, provided by Malgouyres et al. (2021), and construct a bilateral measure on internet access which varies across time due to pairs geographical coverage and the evolving internet speed. Our findings remain consistent even after controlling for this factor.

To further address for endogenenity concerns, we draw inspiration from research that integrates first-nature geographic elements like average elevation and slope to formulate least-cost paths, as in studies such as Faber (2014) and Berger (2019) who aim at addressing endogeneity issues of railroad connections to explain train.[4] We introduce a waterways connection index that interacts with the year. We justify this approach by noting the resemblance between navigable waterway and high-speed rail networks. We do not employ waterway connection as an instrument, as it has a direct influence on co-patenting. Instead, we include it as a control variable. The effect of travel time remains highly significant.

We further test the robustness of our results, we examine lead and lag effects of reduced travel time, and account for what we term *inter-regionalization effects* – these are time trends that drive individuals to seek knowledge sources beyond their regional borders.[5] These trends can be attributed to advancements in telecommunication technologies or the result of best

[4] Another strand of the literature follows the historical route IV approach developed by Duranton and Turner (2012). Hanley et al. (2022) and Li et al. (2022) chose historical rail connections as instruments for the presence of high-speed railways today. The studies by Li et al. (2022) and Tsiachtsiras et al. (2022), HSR connections were instrumented using the postal routes of the Yuan Dynasty. However, in our specific case, this approach is not suitable, as the development of railways in France initially had a more localized focus, primarily serving transportation needs between major cities and their surrounding areas before gradually expanding to encompass long-distance routes.

[5] The identification of inter-regionalization is inspired by the trade literature, particularly in the context of gravity equation estimations, where the effects of globalization must be carefully controlled for as claimed by Bergstrand et al. (2015) and Yotov et al. (2016).

practices established by companies that recognize the importance of combining knowledge from various sources.

In the second part of the analysis, we investigate the heterogenous impact of travel time reduction on (1) different HSR connection intensity (2) different distance thresholds, (3) different groups of regions, and (4) different technology sectors. We find that travel time reduction had more effect on indirectly connected pairs, those that do not have HSR station at both ends but still benefits from an HSR route in the between. We also find that the effect of travel time reduction is higher for long-distance pairs, at more than 200 kilometers, and for pairs of core regions, characterized by higher innovative activity compared to the periphery.

We observe not only an increase in the elasticity coefficient of travel time as regions become more distant but also a more pronounced impact of the predicted rise in co-patents on the observed co-patent growth. In the context of pairs of core regions, and considering various distance thresholds (below 200 kilometers, between 200 and 400 kilometers, and above 400 kilometers), the predicted increase in co-patents, subsequent to reduced travel time, contributes to 0.5%, 9%, and 20.7% of the observed co-patent growth in core regions, within the respective distance ranges.

In the third part of the analysis, we delve into the evolving nature of collaborations. We analyze the impact of travel time reduction on creating new collaborations, maintaining existing ones, and assess its effect on collaboration quality based on citations, invention scope, and technological diversity. Our findings indicate that the HSR network has a dual impact: it fosters the establishment of new collaborations and enhances the sustainability of existing ones. Although it does not directly affect the quality of inventions, as measured by forward citations, it does contribute to broadening the scope of novelty and the multidisciplinary nature of inventions.

In the last and fourth part of the analysis, we want to understand who is the HSR connecting to gain a deeper understanding of the mechanisms underlying our findings. Thus, we quantify the number of co-patents involving star inventors, pairs consisting of a star and a non-star inventor, and pairs comprising two non-star inventors. Here, a star inventor is identified based on their productivity, determined by the number of forward citations received for their previous patents contributions (Akcigit et al., 2018). To establish a connection between our findings and those of Catalini et al. (2020), we also quantify the number of co-patents where inventor in region i collaborate with inventor in region j who has a higher (and lower) productivity than the average productivity in region i.

Our analysis reveals that reduced travel time has a connecting effect across all types of inventors. However, it is noteworthy that pairs involving star inventors exhibit greater sensitivity to travel time reductions compared to other combinations. Moreover, our findings align with the model and results presented in Catalini et al. (2020), where inventors demonstrate a heightened propensity to collaborate with those possessing productivity levels above the local pool's average.

The findings of this study underscore the transformative potential of high-speed railways as investments in driving innovation. These railways serve as conduits, particularly in core regions, for inventors to collaborate, share knowledge, and diversify their inventive endeavors. This emphasizes the need for strategic investments in transportation infrastructure to connect innovation hubs efficiently. Policymakers should take heed of these results and view high-speed railways not just as transportation networks but as innovation enablers. By nurturing and connecting innovation hubs, policymakers can unlock new opportunities for thriving

innovation ecosystem.

The study raises concerns about the growing divide between core and peripheral areas due to the limited benefits of the HSR network for the latter. Core regions, already rich in innovation, further bolster their innovative capacities through collaboration, while peripheral areas, with weaker innovation orientations, do not experience similar gains. In particular, we have no evidence of enhanced collaboration with core regions.[6]

It is plausible that the positive effect arise over an extended period following HSR implementation, aligning with findings for development convergence from economic geography literature. Initially, core regions reap the benefits, with peripheral areas potentially benefiting later, particularly as more infrastructure improvements are made. However, this outcome may be influenced by a potential bias introduced by the use of patent data to investigate innovation collaboration. Indeed, as previously discussed by Shearmur (2012), patented innovations are less prevalent in non-urban environments.[7]

The paper is structured as follows. Section 4.2 introduces the data used in addressing the research question, beginning with patent data (subsection 2.2.1), followed by the construction of novel travel time data (subsection 2.2.2). Section 2.3 presents the conceptual framework and engages in a discussion that aligns with the existing literature. Section 2.4 outlines the empirical framework and details the identification strategy. Section 4.4 unveils the results of the analysis. Finally, section 3.7 concludes.

2.2 Data

This section introduces the data employed in our study, comprised of two primary datasets. To examine the collaborative interactions among inventors, we leverage patent data. Conversely, for assessing the influence of the high-speed rail network, we construct a dataset detailing travel times between major cities within each NUTS3 region in France.

2.2.1 Collaborative Patenting

Data Source

In order to identify innovation collaboration between regions, we use patent data from Morrison et al. (2017). Their data gathers patent information found in the European Patent Office (EPO), under the Patent Cooperation Treaty (PCT), and in the US Patent and Trademark Office (USPTO). Notably, this dataset provides precise geographical location details for patent owners and inventors, down to the most granular level (geocodes are available). Moreover, the identity of inventors have been disambiguated across time and location. The data includes comprehensive information about patent applications, their proprietors and inventors, application and patent grant years, distinct claims, technological domains of the inventions, and citation data.

[6] This issue is underscored by the findings of De Noni et al. (2018), indicating that involving external inventors, particularly those from knowledge-intensive regions, has a more pronounced positive effect on sustaining higher regional innovation performance.

[7] Eder (2019) provides a critical survey and research agenda on the subject of innovation in the periphery.

Sample Selection

Patents owned by applicants to EPO localized in France. From the vast array of 9,290,268 patents in the raw dataset, we narrow our focus to a particular set: patents submitted exclusively to the European Patent Office (EPO), which aligns with our interest in French patents. The data lacks information about the French patent office, but many French inventors and owners turn to the EPO to protect their ideas in the broader European market, including France. This refined dataset comprises a total of 2,684,761 patents, spanning the years from 1978 to 2014. We overlap the geographical data of assignees onto a world map. Within the dataset, we identify 190,522 French patents, i.e. owned by an applicant localized in France. Note that applicant and owner are used interchangeably in the text. Among these, a notable 93.02% are exclusively owned by French applicants. To refine our analysis, we narrow our focus to this subset of patents, thereby excluding international co-ownership. Consequently, our dataset is streamlined to comprise 177,493 patents.

Years selection. The establishment of the European Patent Office (EPO) in 1978 resulted in a notable surge in the number of patent submissions during its initial years. However, it is important to note that these patents may not have been newly created during that period. Instead, they could have been previously developed and subsequently submitted to the EPO to safeguard inventions in the European market. Additionally, there is a noticeable decline in patent numbers from 2010 to 2014, possibly due to limited data availability during the data collection made by the authors. As a result, we narrow down our sample to begin from 1980 as the starting year and conclude in 2010 as the final year, which results in a total of 165,020 patents, owned by 27,842 different applicants. Figure 2.7 shows the evolution of the total amount of inventions submitted at EPO from 1978 to 2014.

Inventors localized in France. We match the patent data with the dataset on inventors. 163,539 French patents are left, involving the contributions of 143,264 inventors. For each patent, we meticulously count the number of inventors composing the patent team. Subsequently, we refine the data to inventors with valid geographic coordinates, which correspond to the inventors' residential address. This entails the exclusion of observations characterized by unlocalized inventors, as indicated by geolocation quality indicators such as "unloc," "low," and "high." Remain 160,326 patents and 137,499 inventors. Then, we overlap the geographic coordinates of inventors on a world map. We find that 9.16% of the patents are developed by foreign inventors only, 3.72% by teams of both French and foreign inventors, and 87.12% by French inventors only.[8] We restrict the sample to those patents that are exclusively undergone within metropolitan France.

Further curation involves eliminating low-quality geolocation observations and patents with multiple NUTS3 regions reported by inventors. At this stage, 122,688 patents and 98,091 inventors remain. Among co-patents involving a minimum of 2 inventors, we retain those where at least 2 inventors have high-quality geolocation. The result is a sample of 112,315 patents and 95,381 inventors. Notably, 89.79% of these patents possess comprehensive geolocation information for the entire inventor team, defined by high-quality geolocation for all inventors within the team. Figure 2.23 shows the summary statistics of the proportion of inventor's location information of high quality within patent teams.

[8] We define "French inventors" as inventors who are located within the borders of France.

Identify intra-regional and inter-regional patents. We face a challenge concerning whether to maintain patents that contain complete information or permit those with incomplete information. On one hand, removing incomplete patents reduces the number of observable collaborations, but it accurately identifies patents developed within a region's borders. On the other hand, preserving incomplete patents increases our ability to observe co-patents between regions. However, it can make it difficult to recognize strict intra-regional collaborations, characterized by inventor collaboration confined to the same region. Indeed, if we were to classify a patent as intra-regional solely based on the presence of all well-localized inventors within the same region, despite the lack of well-localized data for some inventors, there would be a risk of overestimating the count of intra-regional co-patents.

If we were to erase those patents with incomplete information, we would encounter another issue. The completeness of a patent is inversely proportional to the number of inventors involved, particularly when it comes to inter-regional co-patents which tend to have more inventors on average.[9] Removing incomplete patents would consequently result in a disproportionate reduction in the number of inter-regional co-patents compared to intra-regional ones.[10]

One one hand, to avoid underestimating observations of inter-regional collaborations, we include both incomplete and complete inter-regional patents in our analysis. On the other hand, to prevent the underestimation of intra-regional patents while implementing a stringent classification approach, we employ the following method: any patent that is collaboratively developed by inventors situated within the same NUTS3 region is considered an intra-regional collaboration, regardless of the accuracy of their geolocation data. However, if a patent in-

[9]To discern any statistically significant distinctions between the average number of inventors within patent teams for intra-regional and inter-regional co-patents, as delineated by the count of NUTS3 regions linked to well-localized inventors, we employ Welch's two-sample t-test. This approach accommodates disparate variances and sample sizes between the compared groups. At the 99 percent confidence level, we reject the null hypothesis positing equal means for the two samples. The 99 percent confidence interval of the difference between inter-regional team size and intra-regional team size is [0.49; 0.54] with sample estimates of 3.21 and 2.70 respectively. The difference is statistically lower than one, which does not actually involve that intra- and inter-regional patents are significantly different in their team size.

[10]To test our hypothesis, we employ a linear probability model with the dependent variable being a dummy variable that equals one for patents with complete information and zero for those with incomplete information. The independent variables include the number of inventors and a dummy variable indicating whether the patent identifies one NUTS3 region or multiple regions, respectively taken values of zero and one. We include year fixed effects to control for any time-related changes in the size of patents' teams, the amount of inter-regional co-patents, and other unobservable factors influencing completeness that may be changing over time. Table 2.24 in the appendix show the results. Column 1 estimates the effect of the team size on the likelihood of information completeness on the set of co-patents. We find that the bigger the patent team, the less likely the patent has complete information. Column 2 estimates the effect of being an inter-regional co-patent on the likelihood of information completeness. Our findings indicate that inter-regional co-patents are 1% less likely to have complete information compared to inter-departmental co-patents, which confirms the previously stated hypothesis. Column 3 includes both independant variables, and because the amount of inventors and the dummy indicating inter-regional collaboration are positively correlated, the coefficient of the later turns out to be positive. For a same amount of inventors within the patent team, inter-regional patents are 5% more likely to have complete information on their inventors. Retaining only complete patents could underestimate intra-regional co-patents with respect to inter-regional co-patents, unlike was expected. We are going to consider this concern by classifying intra-regional patents according to another condition than the completeness of its information. It is worth noting that the R squared is equal to zero in column 2, and does not increase in column 3, meaning that being an intra- or inter-regional patent does not explain variations in the completeness of information.

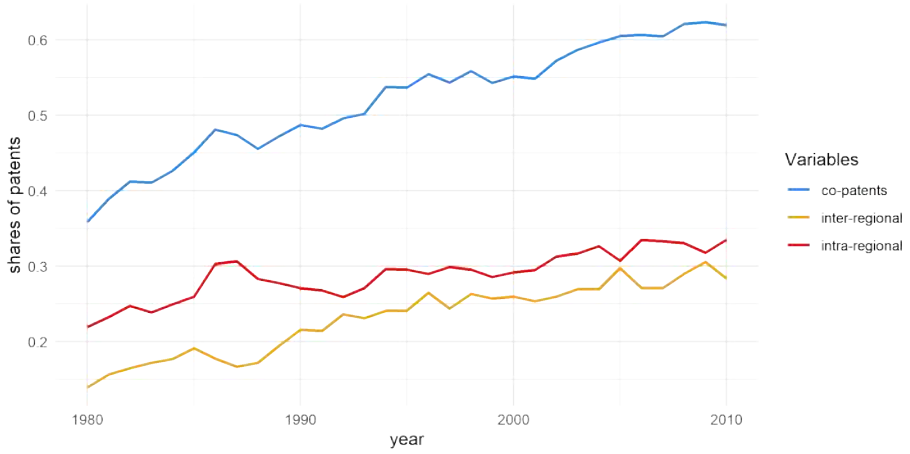


Figure 2.1: Co-patenting within and beyong NUTS3 region borders (1980-2010)

volves an inventor with imprecise geolocation information, placing them in a region different from the clearly identified regions of their collaborators, that particular patent is excluded from the sample. By employing this strategy, we are able to retain an additional 4,268 patents and 2,966 more inventors in our analysis compared to the alternative approach, which would have exclusively considered patents with complete information for intra-regional patent classification.

Therefore, 110,958 patents and 94,448 inventors remain in the sample, where 43.18% are patents developed by single inventors, 28.58% by inter-regional collaboration and 28.24% by intra-regional collaboration. Figure 2.1 depicts the temporal progression of the proportion of patents resulting from collaborative efforts among inventors, relative to all patents included in the sample. Additionally, it portrays the distribution between intra-regional and inter-regional patents.

Final patent dataset. Given that our research focuses on collaboration between pairs of regions, we narrow down our sample to exclusively encompass collaborative patent applications involving a minimum of two inventors. This results in a final dataset of 63,041 patents developed by 77,345 inventors. Notably, half of these patents (50.30%) resulted from inter-regional collaboration. We utilize this set of patents to derive the variables that will form the basis of our analysis, enabling us to depict and elucidate the characteristics and factors underlying collaborative patenting endeavors.

Dependant Variables Computation

We introduce a variety of dependent variables that we intend to examine in relation to the reduction of travel time within region-pair observations.

Number of co-patents within region-pairs. To study co-patenting between regions, we identify each unique pair of inventors working for a same patent. Then, we aggregate the number of unique patents under collaboration for each pair of NUTS3 regions, using the residential location of inventors. Two main approaches can be utilized: the *fully count approach*, as used by Hoekman et al. (2009) and Morescalchi et al. (2015), and the *fractional count approach*, as used by Picci (2010). The choice between the two approaches depends on research

objectives. The fully count approach considers the absolute number of co-patents between regions, offering insights into overall collaboration levels and identifying regions with high co-patenting activity. On the other hand, the fractional count approach allows for relative comparisons by considering the co-occurrence of regions in patents and dividing it by the total number of interactions or inventors. This approach accounts for differences in collaboration intensity, providing a more nuanced understanding of the strength of collaborative relationships between regions. We use both in our analysis.

New versus old collaborations. To attain a deeper insight into the evolution of regional interactions, our exploration hinges on the dynamics of co-patenting, discerning between newly established and pre-existing partnerships among inventors. This analysis leverages the inventors' name and location disambiguation made by Morrison et al. (2017). We identify pairs of inventors within each patent and classify their interactions into two categories: *new collaboration*, which involve inventors who haven't previously collaborated, and *established collaboration*, which already has occured in prior patents. The goal is to ascertain whether the implementation of high-speed railways has predominantly created new interactions between inventors or helped perpetuating existing ones.

Quality of co-patents. In order to assess the quality of collaborative efforts, we leverage three distinct metrics linked to patent quality. The first metric revolves around *forward citations*. This measure quantifies the number of times a patent has been referenced in subsequent patents. A patent's influence on propelling innovation is directly correlated with its citation count, reflecting its capacity to introduce fresh knowledge that aids in advancing innovation. Citations play a pivotal role in enhancing the pool of knowledge as they depend on the insights embedded within the cited patents. In this vein, we assign weights to co-patents based on their amount of *3-year* and *5-year forward citations*, following Akcigit et al. (2018).

Another valuable metric that offers insights into a patent's quality is the quantity of *claims*, following Lanjouw and Schankerman (2004). In the context of patents, a claim is a precise and detailed statement within a patent application that outlines the legal scope and technical features of the invention. These claims define the unique aspects that differentiate the invention from existing technologies and establish the boundaries of the patent owner's exclusive rights. For each region-pair-year combination, we calculate the number of co-patents, assigning weights based on their respective claim volumes. This approach provides a quantitative gauge of the novelty and scope of knowledge encapsulated within the co-patents.

The third and final indicator that we employ to assess the quality of a patent pertains to its *technological fields*, which are associated with each claim (Lerner, 1994). We extract specific information from the 35 subfields encompassed by the International Patent Classification (IPC35) assigned to each patents. Subsequently, we quantify the number of co-patents for each region-pair-year observation, incorporating weights based on the count of distinct technological fields the patent enters in. This metric provides valuable insights into the breadth of technological diversity encapsulated within the collaborative patents. It allows us to gauge the extent to which the co-patents span across various specialized domains, highlighting the innovation's reach and the interdisciplinary nature of the collaborations.

Intra- versus inter-firm co-patenting. Exploring the dynamics of both *intra-firm* and *inter-firm collaboration* is of paramount significance due to the fundamental role that firm bound-

aries play in the diffusion of knowledge. The literature highlights the critical distinction between knowledge flows within firms' boundaries and those that extend beyond. Within firms, knowledge tends to circulate more readily, fostering innovation and facilitating efficient sharing of expertise. In contrast, knowledge dissemination across firms' boundaries is typically more constrained, resulting in a decreased flow of information.

Our dataset, unfortunately, does not permit a granular examination of inter-firm collaborations over time, largely due to the challenge of precisely disambiguating firm names - an inherent limitation frequently encountered when working with patent data. To gain insight into whether collaborations involve single firms or multiple entities, we adopt an approach based on the number of owners associated with each patent. Specifically, we quantify the count of owners linked to each patent and classify those with a single owner as instances of *intra-firm collaboration*. Conversely, patents featuring two or more owners are categorized as instances of *inter-firm collaboration*. Subsequently, we calculate the quantity of co-patents for each region-pair-year observation, distinguishing between the two distinct categories of firm collaboration.

Inventors productivity. To provide a deeper understanding from an inventor-level perspective, our analysis delves into the extent to which collaboration decisions are shaped by the productivity of potential partners. It is particularly relevant in the context of inter-regional collaboration, where the anticipated gains from patents need to surpass the associated travel costs and collaboration expenses for the patent owner to find the partnership worthwhile. To answer the question, we create indicators that aim at representing the productivity of inventors.

We adopt the definition of inventors' productivity as proposed by Akcigit et al. (2018). This measure considers the number of patents each inventor has participated in up to the present year, weighted by the quantity of 5-year forward citations received by each patent. This combined metric, called cumulative productivity, effectively captures both the quantity and quality of an inventor's innovative contributions in the past and current years. The measure grows with time as it sums the additional productivity gained each year, emphasizing the impact of experience on productivity. We begin by compiling a comprehensive dataset that includes all patents from 1978 to 2010, involving inventors present in the inventor-pair dataset. Next, we link the 5-year forward citations to the patents in which these inventors have participated.[11] This enables us to compute the productivity measure for each inventor a at time t according to the forward citations associated to the patents p he has participated in. The measure is

[11]To address the absence of 5-year forward citations data for the final year in our sample, we employ patent information from the years 1978 to 2005. By doing so, we ensure a complete measure of 5-year forward citations for patents in the year 2005. However, we cannot do so for years after 2005. On average, each inventor has participated in 2.24 patents, with a maximum of 156 patents. Considering the weight of their participation, calculated by summing the inverse of the number of inventors on each patent across all their patents, we find that inventors have generated an equivalent of 0.78 patents on average, and 67.17 as the maximum. When accounting for the weight of their 5-year forward citations, we observe that inventors' average productivity is about 0.68, with a maximum of 42.78. Identifying their maximum productivity value over the years, the average reaches 0.84, and the maximum 81. These average figures are very low, but maximum values are shown to be high, showing great heterogeneity across inventors in the data. 58.47% of all inventors have a maximum productivity of zero, meaning that that they never participated to a patent that has been cited in the 5 years following the patent application.

computed as follows:

$$\text{Productivity}_{at} = \sum_{s=t_0}^{t} \sum_{p(a)} \text{citations}_{p(a)t}^{\text{5-year forward}} \tag{2.1}$$

We compute the average productivity at the regional level for each year, utilizing it as a benchmark to derive two indicators. The first indicator determines whether an inventor's productivity at time t exceeds their region's average for that particular year, labeled as *Inventors' productivity above their region's average*. The second indicator assesses whether an inventor's productivity at time t surpasses the average productivity of their collaborator's region, referred to as *Inventors' productivity above the average in their partner's region*. The latter indicator will enable us to assess the hypothesis put forth by Catalini et al. (2020), suggesting that in order to outweigh the costs associated with distance, inventors engage in long-distance collaboration when their partner's productivity exceeds the average productivity of inventors within their local pool of potential collaborators.

Hence, we can quantify the number of co-patents that include two inventors, only one or none with a *productivity above their respective region's average*, as well as the co-patents comprising inventors with *productivity levels above their partner's region's average* or *below*. It enables us to study collaboration dynamics based on inventors' productivity and assess their importance in shaping inventive interactions across regions.

Inventors' set of knowledge: similar VS complementary. We delve into the field of specialty of inventors, creating additional indicators to gain insights into whether inventors collaborate based on a similar or complementary knowledge. Both have their respective advantages in driving innovation. When inventors possess similar expertise, it creates an environment conducive to knowledge exchange, synergy, and smoother communication. This shared understanding fosters increased rapport among collaborators and can enhance problem-solving capabilities. On the other hand, when inventors have complementary knowledge, it can lead to the development of more creative inventions which benefit from a multidisciplinary approach to push the frontier of knowledge. In both cases, the outcome can be productive, allowing inventors to effectively combine diverse skills and innovative ideas to achieve inventions that stand at the forefront of knowledge and technological advancement.

To evaluate the level of technological similarity between collaborators at time t, we employ vectors that capture their degree of specialization with respect to the IPC35 classification of patents they have been involved in the past, i.e. from their first appearance in the dataset t_0 to $t-1$. For each inventor-patent, we ascertain the count of claims associated with their patent and the count of claims within each IPC35 technology they are associated with. By calculating the ratio of the latter to the former, we obtain the proportion of their technological activity among the 35 distinct technology classes for each patent, yielding values ranging from zero to one.

We construct inventor-year vectors to represent technological activity and calculate the cosine similarity index for each collaboration-year. This index measures the similarity between inventors' past specialization, considering their history up to year $t-1$. To compute the similarity index for inventors a and b at time t, we sum their patent-level vectors from their complete history up to year $t-1$. If either inventor has no previous patent participation, we consider their contemporaneous patents, excluding patents where they collaborate. If there are no such patents, we use the technology of their collaboration patent as a last resort.

The technological similarity between inventors a and b is computed as follows:

$$\text{Technological similarity}_{abt} = (A_t \cdot B_t)/(||A_t|| * ||B_t||) \tag{2.2}$$

where A_t and B_t are the technological vectors of inventors a and b, respectively, at time t, including their patents' history from time t_0 to $t-1$. The dot product $A_t \cdot B_t$ represents the similarity between the two vectors, and $||A_t||$ and $||B_t||$ represent their magnitudes (Euclidean norms). The resulting cosine similarity index ranges from 0 to 1,[12] with 1 indicating perfect similarity and 0 no similarity between the vectors.

The median is about 0.96 and testify of a significant level of similarity among most collaborators. Among all pairs-year observations, 27.58% exhibit perfect technological similarity with an index value of one, indicating a strong alignment, while 8.60% show no similarity with an index value of zero. We establish various thresholds to categorize similarity and complementarity. We refer to *similar knowledge* if the cosine similarity index between two inventors is close to 1 (e.g., ≥ 0.8), and to *complementary knowledge* if the cosine similarity index value is below 0.8, implying that the inventors' patent portfolios cover distinct but related areas. These categorizations enable us to count the number of co-patents involving collaborations between inventors with either similar or complementary knowledge.

Independant Variables Computation

The patent data also allow us to compute variables that could explain collaboration formation and maintenance between individuals, which would in turn explain the collaboration trends between regions.

Regional technological proximity. A first candidate is the similarity in the technological knowledge within regions' boundaries (Jaffe, 1986). When regions share similar technological profiles, collaboration is more probable as their expertise and industrial specialization align. Comparable technological profiles imply shared innovation focus, fostering effective communication, mutual understanding, and potential collaboration synergies. To compute the similarity index, we employ the same methodology as with inventors, but at the regional level. We build past technology vectors that quantify patent numbers within each IPC35 category, weighted by the corresponding claim counts. In cases where a region, denoted as i, lacks a past technology vector, we resort to its present technology vector while excluding collaborations with region j. If the latter is also unavailable, we consider the current patents arising from the collaboration between regions i and j.

Inventors' network proximity (bridges). Another candidate involves examining the entire network of inventor collaborations. Shared collaborative networks that extend beyond the direct interactions of two inventors can create an environment where they are more likely to collaborate. If two inventors have common collaborators, it may create the opportunity that they hear from each other and facilitate communication or even foster trust between them, creating a conducive environment for collaboration to emerge between the two inventors. This perspective allows us to uncover the potential influence of shared collaborative networks on co-patenting patterns between regions, as well as reduce potential bias in the estimate of

[12] Cosine similarity index usually ranges from -1 to 1. In our case, patent counts cannot take negative value, so do our cosine similarity index.

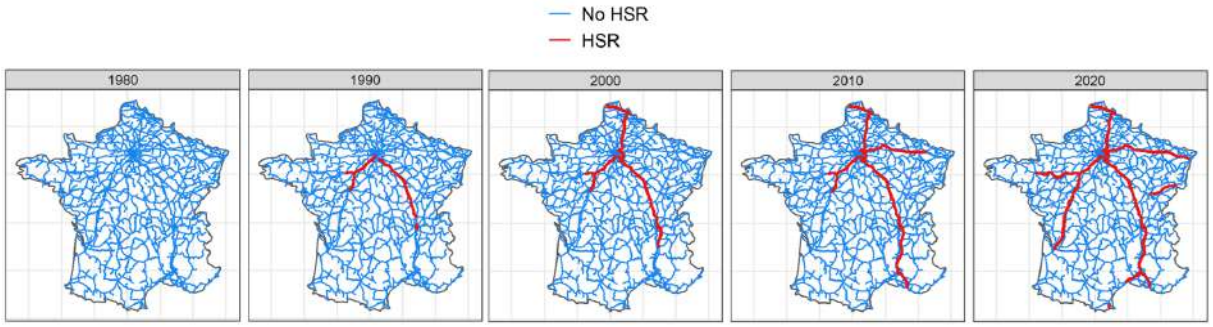


Figure 2.2: 40 years of high-speed railways deployment in France

the effect of travel time, providing a deeper understanding of the dynamics driving inventive collaborations.

Creating an index summarizing two regions' network proximity entails counting the common collaborators shared by inventors a and b, localized in regions i and j respectively, across a range of patents beyond those involving the two individuals (Bergé, 2015). This calculation is performed for both past and current collaborations among common collaborators. To execute this process, we take the following steps. First, (1) we consider patents in which inventors a and b collaborate, spanning from their first collaboration until time t. Second, (2) we compile a list of unique collaborators for inventor a during the same period, excluding collaborations involving inventors a and b - represented by the list mentionned in 1. Similarly, (3) we create a list of unique collaborators for inventor b, excluding collaborations involving inventors a and b. Then, (4) we count the common collaborators, called *bridges*, by identifying the intersection between the vectors of collaborators for inventors a and b. Finally, (5) we associate inventors' regions of residence within each region-pair-year and then sum the count of common collaborators for each respective region-pair-year.

2.2.2 High-Speed Railways

High-speed rail (HSR) is a rail passenger service that operates at significantly higher speeds compared to traditional trains, typically ranging from 200 km/h to 350 km/h. The primary objective of HSR systems is to enhance accessibility and connectivity among well-developed regional urban networks, thereby facilitating mobility for workers and tourists. Serving as a rapid mode of transportation, high-speed railway represents an attractive alternative to air travel, offering passengers reduced travel times, improved efficiency in their journeys and mitigated environmental impact.

High-Speed Railways History in France

France inaugurated its first high-speed railway in 1981, linking Lyon and Saint Florentin, as an initial step towards connecting the capital city of Paris with Lyon, a key French urban center. Over the years, the high-speed rail network has undergone substantial expansion, progressively linking Paris to various major cities: Le Mans (1989), Tours (1990), Lille (1993), Marseille (2001), Montpellier (2001), Strasbourg (2007-2016), Bordeaux (2017), and Rennes (2017). By the close of 2017, the high-speed rail network extended for more than 1,500 kilometers. Given that our patent dataset encompasses the timeframe from 1980 to 2010, our focus in this paper cen-

ters on the network's development within this temporal scope. For further comprehensive information, refer to the online appendix.

Data Source

France's national state-owned railways company, Société Nationale des Chemins de Fer (SNCF), provides schedule datasets containing departure and arrival times for various train services, categorized as TGV (high-speed trains), Intercités (intercity trains), TER (regional express trains), and Transilien (local trains). TGVs operate on both high-speed and normal railways, while the others run solely on normal lines. We downloaded the different datasets in December 2021, which encompass trips made by these trains from the 8th to the 16th of December 2021. The dataset consists of 16 distinct datasets presenting a comprehensive schedule of train operations across metropolitan France. It encompasses information concerning routes, trips, stations, and the departure and arrival times of trains, encompassing all four distinct train types.

Travel Time Computation

Contemporaneous observed travel time within non-stop station pairs. Utilizing the comprehensive train operations schedule outlined earlier, we generate a dataset focused on station pairs through a self-merge process based on route and trip IDs. This procedure enables the extraction of station pairs for every trip. Our focus is on direct connections, where a train completes the journey without any intermediate stops. We capture the departure time from the initial station and the arrival time at the final destination. Then, we compute the observed travel times between connected stations by calculating the time difference between the departure and arrival time. To ensure the uniqueness of information, we retain the minimum travel time value for each non-stop station pair.

Past values of travel time. To estimate past travel time values for pairs where an HSR lies between them today, we rely on three information: (1) the opening of HSR, (2) the contemporaneous average speed of inter-regional trains, i.e. Intercités, since it is comparable to TGV running on normal lines, and (3) the geodesic distance between each non-stop stations pair, which is used as an approximation for the rail distance. For stations pairs with a high-speed railway in the between in December 2021, we uses the contemporaneous observed travel time after an high-speed railways (HSR) has been introduced and we estimate the travel time before the HSR roll-out using contemporaneous Intercités average speed and station pairs' distance. For all other stations pairs, we use the contemporaneous observed travel time.

To compute the estimations of past values of travel time, we estimate the average effect of distance on travel time between stations. We estimate the following regression:

$$\text{travel time}_{od\tau} = \alpha_{0\tau} + \alpha_{1\tau}\text{distance}_{od\tau} + u_{od\tau} \tag{2.3}$$

with *od* the non-stop pair between the origin and the destination stations and τ the train type. Travel time is expressed in minutes and distance is expressed in kilometers.

We find that TGV operating on HSR display a lower coefficient than the other trains. For 10 additional kilometers between two stations, the TGV travel time is expected to increase by 3 minutes, while the Intercités travel time is expected to increase by 5 minutes in average. The same result applies for TGV running on normal lines. From these coefficients, we compute

the implied average cruise speed as follows: $\widehat{\text{speed}}_\tau = 60/\hat{\alpha}_{1\tau}$. Intercités are found to run at an average speed about 111km/h, while we find an average speed of 229km/h for the TGV operating on HSR. Results are presented in the online apprendix (table 4).

Building on these findings, we augment the non-stop station pairs dataset by incorporating the estimated travel time between pairs before the introduction of a HSR connection. The estimated travel time is calculated using the following equation: $\widehat{\text{travel time}}_{od,\text{before HSR}} = \widehat{\alpha_{0,\text{intercité}}} + \widehat{\alpha_{1,\text{intercité}}} \times \text{distance}_{od}$. Here, the coefficients are derived from the Intercités category, with $\widehat{\alpha_{0,\text{intercité}}} = 6.54$ and $\widehat{\alpha_{1,\text{intercité}}} = 0.54$. This approach allows us to estimate the travel time that existed prior to the implementation of high-speed rails. For pairs lacking an HSR connection between them as of December 2021, the past travel time value corresponds to the travel time value of that specific timeframe.

Intra-city correspondance. To ensure a comprehensive train network that accommodates station changes within cities, we also address intra-city travel times. Notably, 11.85% of cities in the dataset feature multiple stations. On average, intra-city distances hover at approximately 4 km, with the third quartile lying at around 6 km. We compute the average speeds of trains running on normal lines for non-stop station connections within a 16 km radius, which represents the maximum distance observed between intra-city stations. We find that Transilien operates at roughly 45 km/h, while TER runs at approximately 67 km/h. Considering the propensity of passengers to utilize buses, taxis, metros, or tramways for station changes – which exhibit speeds akin to that of Transilien – we find it appropriate to adopt Transilien's speed as a suitable benchmark. We compute travel times between intra-city stations as follows: $\widehat{\text{travel time}}_{od,\text{intra-city}} = \overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}} \times \text{distance}_{od}$. Here, $\overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}}$ denotes the average speed observed for Transilien non-stop station pairs within a 16 km range[13]. Incorporating these new intra-city pairs and estimated travel times enriches our schedule dataset.

Shortest path between every stations. Utilizing the travel time data for non-stop station connections, we employ the Dijkstra algorithm (Dijkstra, 1959) to determine the shortest travel time path between every pair of stations within France for the year 2020. Subsequently, retracing our steps back in time prior to the introduction of high-speed lines between two stations, we substitute the high-speed travel time with the estimated travel time obtained through the coefficients from regression 3.1 and the inter-station distance. In essence, this entails replacing high-speed travel time with regular (Intercités) speed for the relevant station pairs in the years preceding the advent of high-speed rail.

The Dijkstra algorithm is iterated 41 times for each year, tracing back from 2020 to 1980, while considering the declining rail speeds. As a result, the dataset showcases travel time changes exclusively linked to the introduction of high-speed railways. Given that some cities possess multiple stations, some city-pairs exhibit more than one travel time value per year. To maintain unique and concise data for each city-pair and year, we retain the minimum travel time for each respective case.

Validation exercise. SNCF has provided a dataset detailing train travel times between a subset of cities in France spanning the years 1920 to 2020. This enables us to compare our estimated travel times with the observed values over the years, allowing for an assessment of

[13] $\overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}} = 0.75$, calculated as 45 km/h divided by 60, given that travel time is measured in minutes.

the accuracy of our dataset. The SNCF subsample comprises 43 distinct city-pairs involving 37 different cities, with many of these pairs featuring Paris. By merging this dataset with our own, we can conduct the following regression:

$$\text{observed travel time}_{odt} = \beta_1 \text{ estimated travel time}_{odt} + \varepsilon_{odt} \tag{2.4}$$

Here, *observed travel time*$_{odt}$ represents the travel time between the origin city o and the destination city d in the year t from the SNCF subsample, while *estimated travel time*$_{odt}$ refers to our computed travel time estimation. The term ε_{odt} accounts for the error term.

The outcomes reveal that our estimated travel time captures approximately 95% of the variance present in the observed travel time within the SNCF subsample. Even when examining travel time variations within city-pairs, the R^2 value remains notably high, at approximately 80%. The effect of the estimated travel time on the observed travel time, denoted as $\hat{\beta}_1$, is approximately 1. This signifies that a one-minute alteration in the estimated travel time corresponds to a roughly one-minute change in the observed travel time. This relationship holds consistent across both overall city pairs and within pairs, whether controlling for city-year effects or not. For more details and visual validation exercises, see the online appendix.

Final dataset. To obtain travel times between NUTS3 regions, we identify the main city of each region based on the highest population size in 1975 from INSEE, the French national institute of statistics, and extract the travel time between these cities. This dataset spans from the year preceding the initial introduction of the high-speed rail system to the present day.

2.2.3 Descriptive Statistics

First Evidence on the Geographical Nature of Collaborations

Amount of regions involved in collaboration. Table 2.1 reveals that patenting collaborations tend to occur primarily between two NUTS3 regions, rather than involving more than two regions. This finding supports the rationale to study the trends of co-patenting activity at the regional pair level. Figure 2.10 provides additional insight into the geographical nature of collaboration, revealing a significantly higher frequency of collaborations within single regions and between pairs of regions compared to collaborations involving multiple regions.

	Min	1st qu.	Median	Mean	3rd qu.	Max
All	1	1	2	1.62	2	10
Among inter-regional patents	2	2	2	2.23	2	10

Table 2.1: Summary statistics on the # of NUTS3 regions involved in collaboration within co-patents

Amount of inventors involved in collaboration. Collaborative patent development typically involves an average of three inventors. It suggests that in cases of inter-regional patents, the most common configuration involves two inventors from the same region, and a third inventor from a second region.

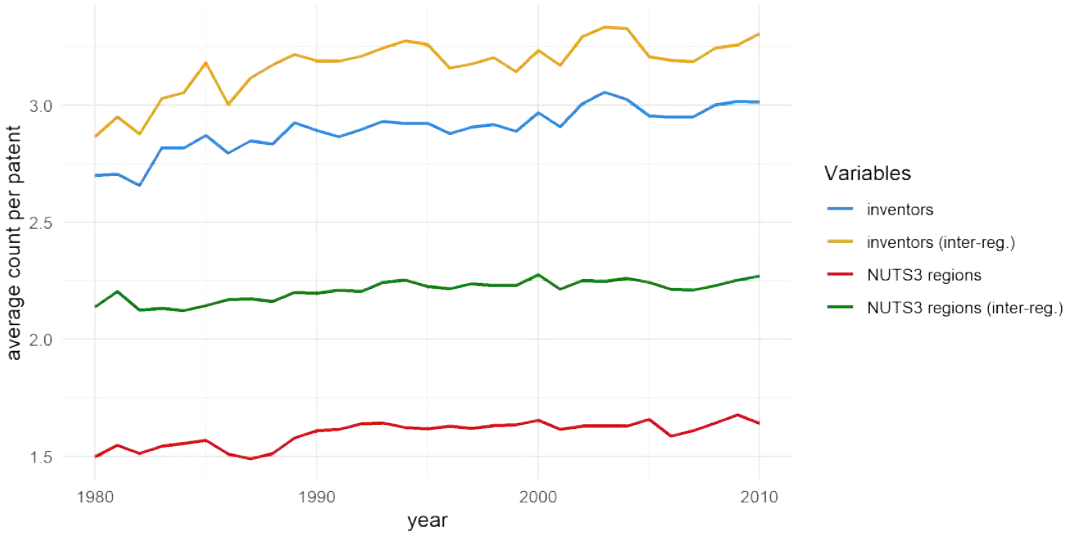


Figure 2.3: Time evolution of average # of inventors and regions within co-patents

Time evolution of the average amount of inventors and regions within co-patents. The figure 2.3 demonstrates the evolution over time of the average number of inventors and regions involved in our sample of patent applications. The graph prominently illustrates a gradual and consistent rise in the average amount of inventors within patent teams, while the average amount of regions involved shows remarkable stability.

Geographical proximity. Statistics on the distance between inventors within patent teams indicate that collaborative innovation is strongly influenced by geographical proximity. Data show that the average distance between inventors is 86 kilometers, while the median distance is only 16 kilometers. It indicates that 50% of collaborations in the sample occur between individuals located within 16 kilometers of each other. For inter-regional co-patents, the median is about 27 kilometers. These findings highlight the fact that inventors are more likely to collaborate with partners who are physically close suggest the importance of face-to-face interactions. The ease of conducting in-person meetings and discussions has been widely acknowledged as a crucial aspect of fostering successful innovation. These face-to-face interactions enable inventors to exchange ideas and information, establish trust, and collaboratively develop novel solutions.

Figure 2.4 depicts the changing trend in the average distance between collaborators over the years for all co-patents as well as for inter-regional co-patents. The average distance between collaborators was approximately 60 kilometers in 1980, which has increased over time and reached more than 90 kilometers in 2010. This suggests that the geographic range of collaborative inventions has become more widespread over the years. In the case of inter-regional co-patents, the average distance between collaborators has increased from 110 kilometers to 140 kilometers, a consistent rise over the same period.

The increase in collaborations across longer distances leads to questions about the impact of improved connectivity between regions, such as the rollout of high-speed railways. These transportation developments enable faster travel over longer distances and fewer stops at stations, which can have a significant impact on reducing average travel times. Such developments may be driving the trend towards longer distance collaborations, and understanding

	Min	1st qu.	Median	Mean	3rd qu.	Max
All	0	6.37	15.51	85.34	44.76	1062.51
Intra-regional	0	2.70	7.12	9.79	13.52	112.03
Inter-regional	0	11.75	26.36	128.69	193.27	1062.51

Table 2.2: Average distance between inventors within co-patent teams

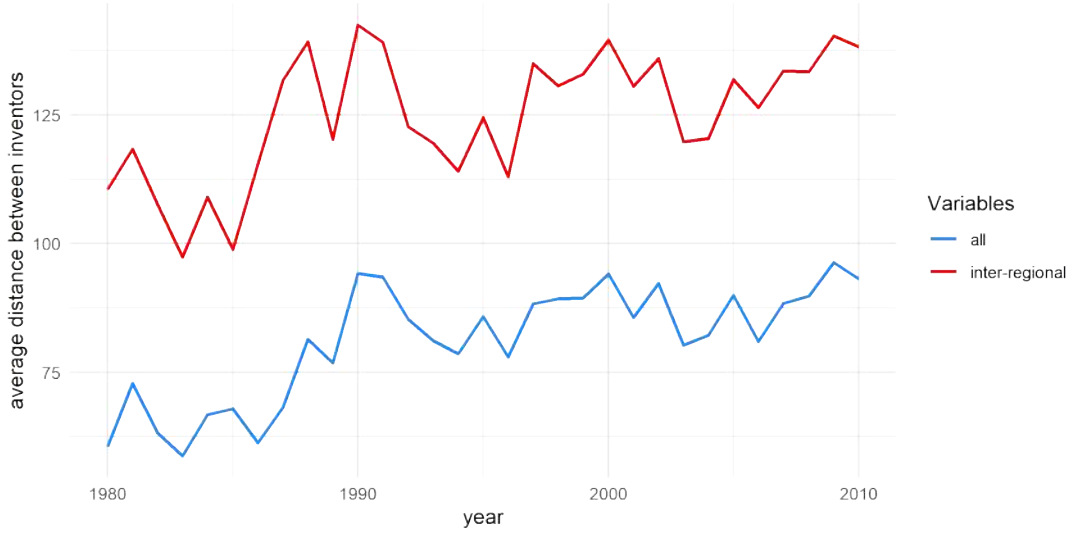


Figure 2.4: Time evolution of average distance between inventors within co-patents

their impact is critical in promoting effective innovation and successful collaborations.

Our paper aims to investigate whether these advancements are contributing to this trend, and whether pairs of regions that are now more connected through faster travel are experiencing a disproportionate increase in collaboration compared to those that did not see a decrease in travel time. By examining the relationship between transportation connectivity and collaboration, this study will shed light on the role of infrastructure in facilitating innovation and collaboration across regions.

Final Dataset

Table 2.3 presents the descriptive statistics on the amount of co-patents developed between NUTS3 regions. Notably, the table is characterized by a prevalence of zero entries across multiple measures, including the minimum, 1st quartile, median, and even the 3rd quartile. Among the entire dataset, the average amount of co-patents developed across all years is about 0.49, indicating a very low collaboration occurrence between regions. The upper limit is established by the maximum count of 435 co-patents, which correspond to a patent that have been developed between inventors of a same region.

Interestingly, the analysis reveals a predominant tendency for innovation to arise within a same region, as indicated by the substantial average of 11.11 co-patents for cases where both inventors originate from the same region ($i = j$). This highlights a distinct intra-regional collaboration trend. Further investigation demonstrates that inventors in geographically contiguous regions display a higher propensity to collaborate in comparison to those at longer

	Min	1Q	Median	Mean	3Q	Max	SD	N pairs	N obs
All sample									
All pairs	0.00	0.00	0.00	0.49	0.00	435.00	6.11	8281	256711
$i = j$	0.00	0.00	3.00	11.11	9.00	435.00	34.25	91	2821
i and j are contiguous	0.00	0.00	0.00	3.53	1.00	291.00	19.64	458	14198
$i = j$ and are contiguous	0.00	0.00	0.00	4.79	2.00	435.00	22.90	549	17019
$i \neq j$	0.00	0.00	0.00	0.38	0.00	291.00	4.84	8190	253890
$i \neq j$ and are not contiguous	0.00	0.00	0.00	0.19	0.00	59.00	1.14	7732	239692
Conditonal on collaborating									
All pairs	1.00	1.00	1.00	4.87	3.00	435.00	18.62	4332	26014
$i = j$	1.00	2.00	5.00	15.20	13.00	435.00	39.28	90	2062
i and j are contiguous	1.00	1.00	2.00	11.53	5.00	291.00	34.17	390	4348
$i = j$ and are contiguous	1.00	1.00	3.00	12.71	7.00	435.00	35.94	480	6410
$i \neq j$	1.00	1.00	1.00	3.98	3.00	291.00	15.28	4242	23952
$i \neq j$ and are not contiguous	1.00	1.00	1.00	2.30	2.00	59.00	3.32	3852	19604
Long distance									
Both in HSR	0.00	0.00	0.00	1.15	1.00	59.00	4.07	168	5208
One in HSR	0.00	0.00	0.00	0.34	0.00	48.00	1.56	2058	63798
None in HSR	0.00	0.00	0.00	0.10	0.00	25.00	0.60	5506	170686
Long distance, conditional on collaborating									
Both in HSR	1.00	1.00	2.00	3.87	4.00	59.00	6.74	156	1542
One in HSR	1.00	1.00	1.00	2.64	3.00	48.00	3.57	1330	8290
None in HSR	1.00	1.00	1.00	1.78	2.00	25.00	1.83	2366	9772

Table 2.3: Summary statistics on # co-patents

distances. This distinction is reflected in the averages of approximately 3.53 co-patents for contiguous regions and 0.19 for regions farther apart, underscoring the role of geographic proximity in facilitating collaboration. Despite the relatively modest average count for co-patents among inter-regional pairs separated by considerable distances, certain instances stand out. For instance, the pair "Paris - Rhône" (Paris - Lyon) achieved a maximum count of 59 co-patents developed in 1999. Note that this pair include the two cities that were initially connected by the first high-speed railway.

The second part of Table 2.3 provides the same statistics conditional on collaborating. This entails analyzing observations where the count of patents is greater than zero. The strikingly low averages below one presented above arise from a significant number of pairs having zero values. In fact, nearly 90% of all observations fall into this category, encompassing approximately 48% of region pairs that consistently exhibit zero values accross all years in the sample. It's worth noting that zero values are more prevalent for long-distance pairs compared to the other types of pairs. Specifically, 95% of the zero values correspond to long-distance pairs (pairs of regions where $i \neq j$ and i and j are not contiguous), while the remaining 5% are observed in contiguous pairs of regions. Similarly, when considering higher averages, the same underlying patterns persist. The prevalence of co-patents being developed within the same region remains notable, with an average count of approximately 15. Moreover, a significant number of co-patents are also observed between contiguous regions, with an average count of around 12, which is notably higher than the count of co-patents between regions located at longer distances, where the average count is about 2.

The final two sections of Table 2.3 delve into the statistics pertaining to pairs of regions

	Average # co-pat in 1980	Average # co-pat in 2010	Average distance (km)	Average growth rate (%)	Amount of pairs N_{ij}
All sample					
All pairs	0.10	0.86	391.84	47.90	8281
$i = j$	3.04	18.38	-	528.27	91
i and j are contiguous	0.72	5.99	85.30	243.89	458
$i \neq j$ and are not contiguous	0.03	0.35	414.28	30.63	7732
Long distance					
Both in HSR	0.13	1.86	412.65	151.19	168
One in HSR	0.06	0.62	419.24	51.93	2058
None in HSR	0.01	0.21	412.47	19.00	5506
Long distance, Both in HSR					
Decrease in travel time	0.14	1.90	421.49	154.32	162
No decrease in travel time	0.00	0.67	-	66.67	6
Long distance, One in HSR					
Decrease in travel time	0.05	0.60	457.42	51.80	1653
No decrease in travel time	0.10	0.70	263.40	52.43	405
Long distance, None in HSR					
Decrease in travel time	0.01	0.21	490.59	19.53	3119
No decrease in travel time	0.01	0.21	310.38	18.31	2387

Table 2.4: Average amount of co-patents in 1980 and 2010, and growth rates

based on the presence of a high-speed rail (HSR) station, unconditional and conditional on collaborating. *Both in HSR* corresponds to pairs where both regions boast an HSR station within their geographical boundaries. *One in HSR* refers to pairs where only one of the two regions possesses an HSR station, while *None in HSR* denotes pairs where neither of the regions has an HSR station. As expected, the former group exhibits higher average and maximum values compared to the second group. Similarly, the second group demonstrates higher values than the latter group.

For a more comprehensive insight into whether the disparities in the average amount of co-patents can be attributed to the HSR network, Table 2.4 provides statistics for the averages in 1980 and 2010 – the initial and concluding years of the dataset – along with the average growth rate between the two years. Not surprisingly, averages are higher in 2010 than in 1980, testifying of the increase in the amount of co-patents developed over time, both locally and at long distances.

While the average growth rate for intra-regional and contiguous pairs is higher compared to longer distances, the latter category exhibits significant heterogeneity, as demonstrated by the divergent average growth rates contingent upon the availability of HSR stations. Notably, pairs connected by HSR stations at both ends exhibit an average growth rate of approximately 150%, while pairs with no HSR exhibit a 19% growth rate on average. Upon delving into the variations within each group – analyzing pairs that underwent reduced travel time against those that remained unchanged – no substantial differences emerge. The forthcoming results of the econometric model are going to provide more insights into the influence of reduced travel time on the significance of co-patenting.

Table 2.5 presents descriptive statistics on travel time and its evolution between 1980 and 2010. A striking observation is the minimal occurrence of travel time reductions among con-

tiguous pairs, implying that the pronounced decrease in travel time is predominantly driven by changes in long-distance relationships. The average travel time reduction hovers around 22% for pairs that have seen an HSR in their travel path, with heterogeneity with respect to the availability of HSR station within the pair. Remarkably, pairs with HSR stations at both ends witness the most substantial decline, averaging around 40%. Meanwhile, pairs with only one HSR station record an average decrease of approximately 26%, while those without any HSR station observe a more modest decrease of around 18%.

Over the whole sample, 3,279 pairs did not see any change in travel time, including the 91 pairs where $i = j$. This subset encompasses about 40% of the total amount of pairs. Those particular pairs will play a pivotal role in our forthcoming econometric model. It will be used as control observations, allowing us to effectively evaluate the impact of travel time reduction on the growth of co-patenting activities.

	Average travel time in 1980	Average travel time in 2010	Average growth rate (%)	Unique pairs N_{ij}
All sample				
All pairs	310.19	261.92	-13.13	8281
Same	-	-	0.00	91
Contiguous pairs	94.42	93.41	-1.11	458
Long-distance pairs	326.28	274.65	-14.00	7732
Conditional on travel time reduction				
All pairs	365.34	285.42	-21.74	5002
Contiguous pairs	111.96	96.02	-17.10	30
Long-distance pairs	366.84	286.55	-21.77	4972
Both in HSR				
Contiguous pairs	48.45	40.20	-14.27	14
Long-distance pairs	281.13	164.63	-39.22	168
Both in HSR, conditional on travel time reduction				
Contiguous pairs	59.86	40.60	-33.29	6
Long-distance pairs	286.85	166.03	-40.67	162
One in HSR				
Contiguous pairs	75.76	73.22	-2.45	98
Long-distance pairs	312.24	238.68	-21.19	2058
One in HSR, conditional on travel time reduction				
Contiguous pairs	104.91	89.41	-15.01	16
Long-distance pairs	338.00	246.64	-26.32	1658
None in HSR				
Contiguous pairs	101.57	101.28	-0.20	346
Long-distance pairs	332.91	291.45	-10.54	5506
None in HSR, conditional on travel time reduction				
Contiguous pairs	165.13	150.80	-9.14	8
Long-distance pairs	386.11	313.71	-18.41	3154

Table 2.5: Average Travel Time and Growth Rate by Group

Finally, Table 2.6 presents additional descriptive statistics for the other dependent variables as described in Section 2.2.1. Specifically, it covers the first 13 variables labeled as *# co-patents*, along with the independent variables introduced in Section 2.2.1, which corresponds to the last two variables in the table.

Variable	Min.	1Q	Median	Mean	3Q	Max.
# co-patents$^{\text{new collab}}$	0.00	0.00	0.00	0.42	0.00	349.00
# co-patents$^{\text{old collab}}$	0.00	0.00	0.00	0.11	0.00	136.00
# co-patents$^{\text{3-year forward citations}}$	0.00	0.00	0.00	0.66	0.00	120.00
# co-patents$^{\text{3-year forward citations}}$	0.00	0.00	0.00	0.27	0.00	634.00
# co-patents$^{\text{claims}}$	0.00	0.00	0.00	2.33	0.00	2680.00
# co-patents$^{\text{technology field}}$	0.00	0.00	0.00	0.79	0.00	676.00
# co-patents$^{\text{intra-firm}}$	0.00	0.00	0.00	0.43	0.00	400.00
# co-patents$^{\text{inter-firm}}$	0.00	0.00	0.00	0.06	0.00	49.00
# co-patents$^{\text{leaders}}$	0.00	0.00	0.00	0.29	0.00	293.00
# co-patents$^{\text{top average productivity}}$	0.00	0.00	0.00	0.11	0.00	124.00
# co-patents$^{\text{catalini}}$	0.00	0.00	0.00	0.16	0.00	189.00
# co-patents$^{\text{similar}}$	0.00	0.00	0.00	0.36	0.00	337.00
# co-patents$^{\text{complementary}}$	0.00	0.00	0.00	0.19	0.00	168.00
Technological similarity	0.00	0.23	0.41	0.40	0.57	1
Bridges	0.00	0.00	0.00	0.20	0.00	754.00

Table 2.6: Summary statistics on the other dependant and independant variables

2.3 Conceptual Framework

Innovation creation is a multifaceted process intricately linked to interactions among inventors, often scientists, engineers, and researchers. This framework delves into the pivotal role of face-to-face interactions and geographical proximity in catalyzing innovation, particularly in patenting activities.

The Power of Face-to-Face Interactions. In the realm of collaborative innovation, inventors come together to collectively develop inventions, with the possibility to apply for a patent if the invention demonstrate novelty and non-obviousness. Patents symbolize the formal recognition and protection of novel ideas, granting the invention's owner exclusive rights over their creations. To seek patent protection, the innovation owner must be motivated to strengthen their intellectual property portfolio and leverage their inventive advances to gain commercial advantage.

A distinctive feature of collaborative innovation is the complex exchange of knowledge among inventors. This exchange involves the fusion of diverse expertise, viewpoints, and technical insights. As inventions become more complex, requiring a diverse fusion of skills, the complexity of knowledge exchange, both in terms of depth and breadth, also intensifies. This complexity calls for a direct, face-to-face interaction model that transcends the limitations of remote communication. As stated by Gaspar and Glaeser (1998), in-person discussions enable a deeper comprehension of intricate technical matters, facilitating real-time brainstorming, immediate clarification, and iterative refinement of ideas. The immediacy of these discussions promotes nuanced exploration, shared understanding, minimizes misinterpretations, and accelerates the resolution of inherent technical challenges in innovation projects. Beyond technical discussions and validation, non-verbal cues, body language, and interpersonal dynamics contribute to a richer understanding of each inventor's unique insights.

Collaborative innovation projects, particularly those aimed at patenting, frequently in-

volve industrial-use inventions in the fields of *Human Necessities*,[14] *Performing Operations - Transporting*,[15] *Chemistry - Metallurgy*,[16] *Textiles - Paper*,[17] *Fixed Constructions*,[18] *Mechanical Engineering - Lighting - Heating - Weapons- Blasting*,[19] *Physics*,[20] and *Electricity*,[21] following the International Patents Classification (IPC). These inventions, often rooted in scientific discoveries or technological advancements, require laboratory work, equipment handling, physical experimentation, prototyping, and validation. Instances of this can be observed in various contexts, such as medical research laboratories, materials testing facilities, and manufacturing plants. These environments require close collaboration among inventors within specialized laboratories equipped with essential tools. The challenge often arises from the impracticality of transporting these specialized equipment to different locations. Face-to-face interactions become pivotal in these scenarios, offering inventors an environment where they can collectively work on intricate laboratory tasks and conduct physical experiments.

Team Composition and Expertise. Innovation projects are often led by a firm or inventor who strategically assembles a team of individuals with specific skills, expertise, and potential for collaboration. The team's composition is meticulously determined to ensure coherency, optimal combination and complementarity of skills and knowledge, trust, and optimal communication among team members. For instance, a study by Cummings and Kiesler (2005) underlines that diverse teams, when appropriately composed, possess an enhanced ability to generate novel ideas and foster creative problem-solving, attributing these outcomes to the amalgamation of distinct viewpoints and cognitive approaches. The presence of a star inventor is also reported as an important component for successful innovation. In particular, Akcigit et al. (2018) models the quality of research teams production which increases in the team leader's productivity and in the team size.

Trust, a foundational element of effective collaboration, is inherently tied to the dynamics of team composition (Miguelez, 2019). In-person interactions offer a unique avenue for the cultivation of trust, as they provide the space for non-verbal cues, shared experiences, and interpersonal connections to develop organically. Empirical studies, such as those conducted by Jarvenpaa and Leidner (1998), emphasize that face-to-face interactions contribute significantly to the establishment of trust, thereby enhancing team cohesion and knowledge sharing. This interpersonal trust forms the bedrock upon which seamless knowledge exchange and collaborative efforts are built, ultimately driving the innovation process forward.

Leveraging External Knowledge. Leveraging external knowledge from diverse geographic origins has emerged as a powerful driver of innovation. Despite the challenges presented

[14] *Human Necessities* field cover a wide range of essential aspects of human life, including healthcare, agriculture, household items, safety, and recreation.

[15] *Performing Operations, Transporting* field relates to processes and methods involved in various operations, as well as inventions related to transportation and vehicles.

[16] *Chemistry, Metallurgy* field refer to patents related to chemistry, chemical processes, and metallurgical inventions.

[17] *Textiles, Paper* field includes patents related to textiles, fabrics, and paper products.

[18] *Fixed Constructions* includes inventions related to construction, buildings, and architectural structures.

[19] *Mechanical Engineering, Lighting, Heating, Weapons, Blasting* field covers a wide range of mechanical inventions, including machinery, lighting technology, heating systems, and weaponry.

[20] *Physics* field refers to patents related to various aspects of physics, including optics, electromagnetism, and acoustics.

[21] *Electricity* field encompasses inventions related to electrical circuits, devices, and technologies.

by expenses like transportation and temporary residence, the potential benefits of integrating varied expertise and perspectives can make this investment highly worthwhile. Akcigit et al. (2018) liken the innovation process to a dance, where partnering with inventors from distant regions enriches projects with unique insights cultivated within their own innovation ecosystems, intuition empirically verified by De Noni et al. (2018). By assimilating this distinct knowledge, projects gain fresh viewpoints, alternative methodologies, and innovative approaches to problem-solving. Furthermore, the authors show in their model that enhanced quality in research projects contributes to improved innovation, and ultimately serves as a driving force behind overall economic growth.

Catalini et al. (2020)'s reinforces the notion that diverse perspectives can lead to transformative outcomes in innovation projects. Their study delves into how travel costs shape collaboration dynamics, emphasizing that reduced travel time between distant localities expands the pool of potential collaborators available to inventors. This heightened accessibility increases the likelihood of finding the ideal co-author for collaborative endeavors. Their empirical inquiry unveils that the introduction of new airline connections, by reducing travel time, encourages collaboration among inventors with elevated expertise. These connections often link inventors who surpass the average competencies within their local talent pools. This pattern substantiates the idea that bringing together inventors across distances is particularly fruitful when the distant partner offers exceptional skillsets. This phenomenon enriches collaborative dynamics and amplifies the potential for innovation.

Moreover, leveraging external knowledge for research projects offers a powerful defense against the lock-in effect, as outlined by Boschma and Lambooy (1999) and Visser and Boschma (2004). This effect, which occurs when projects become deeply rooted within existing knowledge, dampens innovation and limits exploration. Incorporating external insights injects diversity and fresh perspectives, breaking free from insularity and enabling dynamic adaptation. Boschma's lock-in concept warns against relying solely on internal knowledge, which can lead to intellectual stagnation. By contrast, external knowledge disrupts this inertia, stimulating new thinking and reevaluation. Moreover, integrating external knowledge aligns with open innovation principles advocated by Chesbrough (2012). It counterbalances the myopic tendencies of lock-in, fostering a culture of continuous learning and evolution.

Synergy of Telecommunication Technologies. While face-to-face interactions play a central role in innovation, telecommunication technologies can offer a valuable complement rather than a substitute as defended by the literature opened by Gaspar and Glaeser (1998). Telecommunication tools enable real-time communication, information sharing, and collaboration across distances, helping to bridge gaps and enhance coordination among geographically dispersed teams. However, they often work in synergy with face-to-face interactions, rather than fully replacing them, as non-verbal cues and deeper knowledge exchange remain challenging through virtual means. As previously advocated, this is especially true in the context of industrial innovation activities, which are the subject of patents applications.

Regional Innovation. As previously discussed, innovation depart from both local and external knowledge sources, often culminating from collaborative projects. In the context of collaboration, the assimilation of external knowledge hinges on a multifaceted interplay, encompassing factors such as the collaborators' expertise and learning abilities, the entailed costs of travel and communication, and the symbiotic benefits arising from shared learning and joint

problem-solving within the collaborative team.

The regional innovation output, symbolized by the variable $RegInnov_i$ in equation 2.5, is captured by the function F. The function takes into account the distinctive traits of local expertise and endowments $InnovCap_i$, encapsulating the inherent knowledge accrued by individuals within the region, as well as factors in the cumulative collaborative efforts, resulting in the synthesis of shared knowledge between inventors from regions i and j, denoted as $Collab_{ij}$ - note that i and j can be equal. This collaborative initiative is further guided by the expertise and corresponding endowments of the collaborative relationship between the two regions, embodied in variables $InnovCap_i$ and $InnovCap_j$, in conjunction with the involved travel and communication costs $ComCost_{ij}$, which play a critical role in knowledge dissemination throughout collaborations.

$$RegInnov_i = F\Big[InnovCap_i, Collab_{ij}(InnovCap_i, InnovCap_j, ComCost_{ij})\Big] \tag{2.5}$$

The nature of collaboration is inherently intertwined with the respective expertise and endowments of the collaborators, $InnovCap_i$ and $InnovCap_j$, as well as the communication costs $ComCost_{ij}$ incurred during the knowledge-sharing process. An inverse relationship between communication costs and collaboration is emphasized due to the potential hindrance posed by elevated costs on effective knowledge exchange, potentially curtailing the full advantages of collaboration. The combined influence of the collaborators' expertise contributes positively to collaborative outcomes, underscoring that heightened expertise or endowments augment the aggregated amount of collaborations between regions.

This function encapsulates the interplay between local and external knowledge sources, highlighting the pivotal roles of collaboration and associated costs in molding the trajectory of innovation. In this paper, our primary focus lies in examining regional interactions through collaboration, as embodied by the variable $Collab_{ij}$. We are particularly intrigued by its responses to shocks in communication costs, notably travel time. Our goal is to evaluate how enhancements in transportation infrastructure, such as high-speed railways, impact innovative regional interactions. The following sections introduces the empirical model.

2.4 Empirical Framework

2.4.1 Gravity Equation

Economic studies have extensively employed gravity models to assess the interaction between two distinct geographic entities, drawing upon the principles of Newton's law of gravitational force. These models establish a relationship between the entities' respective mass and the distance separating them, serving as a framework for estimating their level of interaction. While initially introduced by Tinbergen (1962) in the context of trade, gravity models have found substantial adaptation within the realms of international trade and migration literature. Nevertheless, the gravity model can be applied to various other contexts involving bilateral interactions between entities.

Recent research studies have adapted the gravity model to explain bilateral collaborations. The first attempt was made by Guellec and van Pottelsberghe de la Potterie (2001). Using patent data, they identify collaborations between inventors at the country level and find that collaboration between inventors are largely explained by geographical proximity, that the

measure by mean of common national border. They also find that internationalisation of a country's technological activities decreases with the level of its GDP and its R&D intensity. In other words, the larger the country, the easier to find collaborators inside the country. Their results suggest that inventor collaborations are substantially localized over close distances. This statement is confirmed by Hoekman et al. (2009) and Frenken et al. (2009).

The literature on collaborations in relation to the gravity equation can be categorized into three distinct tiers: analysis at the national level, exemplified by studies such as Picci (2010) and Montobbio and Sterzi (2013); examination at the regional level within the European context, as demonstrated by research from Hoekman et al. (2009), Morescalchi et al. (2015), and Tóth et al. (2021); and regional-level investigations within a single nation, with von Proff and Brenner (2014) focusing on Germany and Catalini et al. (2020) on the United States. A considerable surge of papers specifically investigating the effect of high-speed railways introduction on inter-city collaborations in the case of China has emerged lately, with studies by Hanley et al. (2022), Yao and Li (2022) and Kang et al. (2023). Across all of these cited papers, a consistent pattern emerges, underscoring a negative impact of physical distance on collaboration among inventors. Moreover, this evidence holds its ground across diverse types of data sources, including scientific publications and patent records. In this paper, we are going to provide a regional-level investigation of the impact of distance, travel time and infrastructure improvement on collaborations in the case of France.

We present a gravity equation that links the co-patenting activity between two regions to their respective innovative capabilities and communication costs. Innovative capabilities encompass various factors such as the number of patents, R&D expenditures, the availability of innovative businesses, academic expertise, and skilled researchers and inventors. On the other hand, communication costs represent the capacity of individuals, localized in two distinct areas, to communicate and exchange knowledge and ideas to conduct an innovative project under collaboration. Communication costs involve transportation costs, relating to distance and travel time, as well as communication technologies and infrastructure such as internet connectivity and teleconferencing tools. Additionally, inter-regional co-patenting is influenced by inventors' network proximity, that can be measured as the amount of common collaborators, as well as by technological proximity, based on shared specialization and knowledge within patenting fields. Network proximity fosters interaction and trust among inventors, while technological proximity enables effective collaboration through shared expertise and jargon.

Equation 4.4 presents the gravity equation of co-patenting.

$$\#\text{ co-patents}_{ijt} = a \times \text{ComCosts}_{ijt}^{\beta} \times \text{InnovCap}_{it}^{\gamma} \times \text{InnovCap}_{jt}^{\delta} \times \eta_{ijt} \tag{2.6}$$

$$\iff \#\text{ co-patents}_{ijt} = \exp[\alpha + \beta \ln \text{ComCosts}_{ijt} + \gamma \ln \text{InnovCap}_{it} + \delta \ln \text{InnovCap}_{it}]\eta_{ijt} \tag{2.7}$$

where i and j denote localities, specifically NUTS3 regions in our context, also referred to as *départements*. Subscript t represents the year, encompassing the range $t = [1980; 2010]$. Coefficients β, γ and δ represent the elasticity of collaboration with respect to each component. Estimating this equation becomes straightforward when transforming it into logarithmic form and again into the exponential form. Doing so allows us to estimate elasticity coefficients, which express the percentage change in the amount of patents developed between regions i and j if the independant variable experience a 1% change. We estimate the regression using a Pseudo-Poisson Maximum Likelihood model (PPML) as recommanded by Silva and Tenreyro

(2006).[22] This method is particularly well-suited for dependent variables that encompass zero values and exhibit a count distribution. It also possesses the ability to effectively manage heteroskedasticity within the dataset, a characteristic that aligns with our co-patenting dataset.

2.4.2 Identification Strategy

Baseline

Structural Gravity. Since we want to reduce the problem of omitted variable due to the lack of data we have on the innovative capacity of regions as well as estimate the effect of travel time reduction on inter-regional collaborative endeavors, we estimate a three-way fixed effects model, as follows:

$$\#\text{ co-patents}_{ijt} = \exp\Big[\rho_{ij} + \gamma_{it} + \delta_{it} + \beta \log(\text{TravelTime}_{ijt})\Big]\eta_{ijt} \tag{2.8}$$

where γ_{it} and δ_{it} are regional fixed effects, controlling for all possible regions' time varying characteristics influencing co-patenting activity between regions i and j, which help reducing the concern of omitted variable bias. They also integrate the information of better accessibility to all other regions following the implementation of HSR, known as multilateral resistance terms. It encompasses the idea that the degree of collaboration intensity between two regions is shaped not solely by their individual innovative potential and the associated communication costs, but also by their relative remoteness in both geographical and innovation terms compared to other regions.

Parameter ρ_{ij} represents a consistent set of dyadic effects that remain static over time. This parameter encompasses various time invarying communication costs and unchanging resistance factors, such as historical relationships, cultural ties, and geographical features between the regions in the pair, such as distance. Its inclusion serves to tackle the issue of endogeneity, arising from the possibility that firms and inventors might decide on their locations based on travel time to their collaborators. This scenario can potentially introduce a challenge of reverse causality in the regression analysis. Moreover, we believe that ρ_{ij} also controls for the likelihood of regions i and j to be connected by HSR.

The extensive set of fixed effects offers a robust method for establishing the causal relationship between co-patenting and travel time. This achievement is made possible exclusively through the variations in travel time between regions i and j, as any unchanging travel time for a specific pair is already accounted for within the pair fixed effects, ρ_{ij}. All pairs with unchanged travel time are going to serve as control observations to effectively estimate the elasticity coefficient β, which represent the impact of travel time reduction between regions i and j on the collaborative development of patents.

Intra-regional co-patenting. As mentioned earlier, the control group for estimating the effect of travel time reduction consists of pairs that do not encounter any changes in travel time. This control group encompasses pairs with identical regions, where $i = j$. To examine the significance of changes in the composition of this control group, we will exclude intra-regional pairs from the sample. Consequently, the remaining pairs in the control group will primarily consist of pairs with contiguous regions. Moreover, these intra-regional pairs will

[22] We employ the R package "fixest," developed by Bergé et al. (2018), to estimate high-dimensional fixed-effects generalized linear models (feglm) with a Poisson link function.

be excluded from the analysis to focus solely on comparing co-patenting patterns among long-distance pairs. This allows us to investigate the impact of travel time reduction specifically for pairs with substantial geographic distances.

The inclusion of intra-regional pairs in the model include the possibility of their influence on collaboration dynamics and potentially introduce deviations from inter-regional collaboration patterns. By excluding them, we can gauge whether the coefficient β is sensitive to this change and whether its magnitude differs significantly when intra-regional pairs are included. If the coefficient β exhibits higher magnitude with intra-regional pairs, it suggests that HSR redirect a significant portion of patenting projects from domestic collaboration towards inter-regional collaboration.

Communication costs

The incorporation of our set of fixed effects doesn't fully account for omitted variables that vary both across and within pairs at the ijt level, and these variables are included within the error terms, which can be a problem for the estimation of β if they are correlated to travel time. We are considering two potential candidates identified in the literature that might influence co-patenting activity.

Inventors' network interconnectivity - bridges. The first candidate emanates from the micro-level and is rooted in the inventors' network. This candidate revolves around their shared collaborators, also referred to as *bridges* in existing literature (Bergé, 2015). We aggregate this measure at the regional level to derive the concept of Regional Network Proximity. This metric reflects the level of interconnectivity between two regions in terms of their inventors' collaborative network. Consequently, higher values indicate greater interconnectedness between regions in their inventors' collaborative networks, which in turn suggests a higher propensity for collaboration.

Technological similarity. Another potential candidate is the Regional Technological Proximity, as computed following Jaffe (1986). This factor incorporates the overlap of various technological fields in which regions specialize, as indicated by the patents they develop within each respective field. When regions possess a higher similarity index in terms of their technological domains, the likelihood of finding mutual interest in collaboration increases. This is due to the shared knowledge and expertise stemming from their similar technological pursuits.

Endogeneity

HSR connectivity. While the inclusion of fixed effects notably enhances the robustness of the identification strategy, it is crucial to recognize that reduction in travel time, implied by the implementation of a HSR connection, could still be impacted by fluctuations in the quantity of co-patents developed. This influence is particularly pertinent for region pairs that are directly linked with an high-speed railway. It is worth considering the potential scenario where an HSR connection is established between two cities explicitly to enhance collaboration, presenting a challenge when estimating the coefficient β. Nonetheless, the influence of the HSR network on travel time is significant, reaching regions that are not directly linked to the network.

To mitigate this potential confounding, we refine our sample to region pairs lacking an HSR station at either the origin or destination. This approach allows us to explore the exogenous impact of travel time reduction on co-patenting among regions indirectly affected by the network due to their fortuitous proximity to cities with HSR stations. Consequently, the coefficient β will signify the percentage effect of a 1% reduction in travel time for pairs without a direct HSR connection.

Lead and Phasing-In Effects. Incorporating lead effects, often referred to as anticipation effects, will allow us to assess whether there is a discernible upward trend in co-patenting before the reduction in travel time. This consideration is crucial for investigating any potential endogeneity concerns in our estimation. It is unlikely that if lead effects exist, they stem from the possibility that certain firms and inventors adapt their collaborative network in anticipation of the forthcoming improvements in accessibility conditions. Furthermore, we will explore the time-varying impact after travel time reduction, commonly referred to as the phasing-in effect. This phenomenon can be attributed to the fact that the effects of travel time reduction might not immediately translate into increased collaborative patents but might take time to manifest.

To address lead effects, we introduce dummy variables that takes the value of one, for three and two years before the first travel time reduction, and zero otherwise. To evaluate the lag effect of travel time reduction, we introduce lags of zero, one, two, and three years, capturing the impact in the years following the first travel time reduction. Additionally, we include a dummy variable representing a period of four years and beyond after the initial travel time reduction. Due to the limited time span of our sample, we refrain from including numerous years preceding and following travel time reduction. This approach is aimed at preventing the omission of observations for certain pairs. Notably, the beginning of our patent data availability predates the implementation of the first railway by one year, and the final railway implementation precedes the last year in our sample by three years. We define the year of first travel time reduction as $t_{0ij} = \min\{t/\Delta^{t,1980}\text{travel time}_{ijt} < 0\}$ over all years t within each region-pair.

We do not expect leading effects to be significantly different from zero since overcoming long distance at high travel time incurs both psychological and time-related costs, which do not favor patenting collaboration. On the other hand, we expect the pashing-in effects to be significant after a certain period following a reduction in travel time, as innovative work typically requires some time to materialize. As time goes by, we can expect these effects to continue growing gradually. However, it is also plausible that the impact of reduced travel time could stabilize or even gradually diminish, potentially influenced by the widespread adoption of high-speed rail connections, resulting in an increased level of regional interconnectedness.

Robustness: Communication Technologies and Inter-Regionalization

Information and Communication Technologies. If we assume that virtual communication between individuals from two different regions evolve differently over time with respect to the pair of regions involved, we must control for the bilateral internet access. We use the data of Malgouyres et al. (2021), which provide the geographic coverage internet broadband staggered roll-out and expansion within municipalities in France from 1995 to 2007, expressed as the proportion of land area covered with values between 0 and 1. We fill missing values

as 0 for the years before 1995 and as prediction resulting from the fit of a time series non-linear model within each municipality, i.e. which fits a logistic function to define the concave evolution of the broadband coverage over time within each municipality.

First, we create an index summarizing the bilateral internet access of regions over time by considering their own broadband coverage. The regional broadband coverage is computed as the weighted sum of broadband coverage over cities within each NUTS3 region, with the population size within each city in 1975 from INSEE as a weight,[23] to take into account the unequal repartion of population within regions.

For each region and year,the broadband coverage is computed as follows:

$$\text{broadband coverage}_{it} = \frac{\sum_{m=1}^{M_i} \text{broadband coverage}_{m,i} \times \text{population size}_{m,i,1975}}{\sum_m \text{population size}_{m,i,1975}} \tag{2.9}$$

where $\text{broadband coverage}_{m,i,t}$ refers to the broadband coverage proportion in municipality m located in region j in year t, $\text{population size}_{m,i,1975}$ refers to the population size of municipality m located in region i in 1975, and M_i is the total amount of cities within regions i.

Then, we compute a bilateral index which is supposed to represent the probability of communication via internet between individuals in regions i and j. We assume that it is directly proportional to the broadband coverage and population size in each region, as well as to the yearly average internet speed.[24] The expected bilateral internet-communication intensity is computed as follows:

$$\text{internet access}_{ijt} = \text{broadband coverage}_{it} \times \text{broadband coverage}_{jt} \times \text{internet speed}_t \tag{2.10}$$

Incorporating this measure into the regression will account for the introduction of the internet and will enable the estimated β to represent the impact of travel time reduction, net of the effect of the growing virtual interconnectedness between regions on collaboration, and the improved capacity for communication that does not require to physically overcome the geographical barriers.

Navigable waterways. In continental France, rivers encompasses around 18,000 km of water routes, out of which 8,500 km are navigable. These navigable routes, including both natural waterways like rivers and man-made ones like canals constructed in the 19th century (such as the famous Canal du Midi), aimed to create an extensive network connecting regions and ensuring efficient transportation across the country. These waterways likely played a pivotal role in establishing and fortifying trade routes, shaping interactions and economic ties between cities. They may have contributed to regional integration and facilitated economic cooperation, with enduring impacts even today.

Interestingly, waterways and the HSR network exhibit comparable geographical network configurations as shown by figure 2.12 in Appendix 2.C. This similarity in their network shapes can be quite striking, even though their modes of transportation and underlying infrastructure are vastly different. Waterways, such as rivers and canals, naturally follow geographical contours and topography, often resulting in interconnected routes that align with natural landscapes. Similarly, HSR networks are also designed with a consideration for the geographical

[23] We consider the population size in 1975 in order to avoid simultaneity problems due to location decision of people.

[24] The data on the yearly average internet speed has been collected from this website We fill the missing years with 0 kbps before 1995 and with linear interpolation for years after 1995.

layout of regions they connect. The alignment of HSR tracks is optimized to minimize curves and gradients, mirroring the contours of the land to ensure a smoother and faster travel experience.

This parallel between waterways and HSR networks highlights the importance of geographic context in designing efficient transportation systems. Just as rivers and water bodies naturally form interconnected paths that facilitate the movement of goods and people, HSR networks are strategically planned to create swift connections between regions, leveraging the natural geographical features to enhance connectivity. Given the historical connections fostered by waterways, collaboration between waterway-connected cities might be more important than for other pairs of cities. This could have influenced the positioning of HSR connections, as historical relationships possibly factored into decisions on rail locations.

To address these historical relationships associated with waterways, we compute a variable that quantifies the distance between region-pairs using the waterways and coastline network. This index assumes zero value otherwise. We intend to incorporate this data into three distinct components. Firstly, we will include a dummy variable encompassing region-pairs not connected by waterways, interacted with a time trend. Secondly, a dummy variable will encompass region-pairs with short-distance waterway connections, also interacted with time trends. Lastly, we'll consider region-pairs connected by long-distance waterways, once again interacted with time trends. The latter is anticipated to exhibit a strong correlation with the presence of an HSR in the between.

Openess to the outside. The estimate of travel time reduction could encompass broader collaboration trends, such as an increasing propensity to be open to the external sources of knowledge, and thus exhibit a potential upward bias that would increase the magnitude of our estimated coefficient, especially when including intra-regional pairs. To tackle this issue we draw inspiration from the approach of Bergstrand et al. (2015) within the context of trade and Regional Trade Agreements (RTAs). Alongside using co-patenting flows within a region as the dependent variable, we introduce a time-varying border variable, denoted as $Same_{ij} \times t$. This variable takes the value of year t if $i = j$ and zero otherwise. The coefficient of $Same_{ij} \times t$ is expected to be negative if, over time, regions increasingly open up to external influences. This coefficient represents this trend as linear in time. If the coefficient beta associated to travel time is still significantly different from zero, we can be confident that co-patenting between regions react to reduction in travel time.

Agglomeration. Similarly, we introduce a variable $Contiguous_{ij} \times t$ in order to assess whether there is some linear time trends in co-patenting between contiguous regions, namely regions that share an administrative border. If contiguous regions progressively engage in more co-patenting collaborations over time, we anticipate the associated coefficient to be positive. Since the departments involved in a co-patent are the residence location of the inventors, the positive coefficient could be attributed to agglomeration patterns. As economic growth takes place in core cities and regions, the population may choose to reside in the surrounding areas where rents and land costs are more affordable, while still commuting to work in the city center.

Death of distance. To robustly identify the impact of travel time reduction, we introduce a time trend by distance, denoted as $log(distance_{ij}) \times t$. This control addresses changes over time in the impact of distance and the costs associated with traveling for collaborative innovation.

This control might be particularly robust, but potentially encompassing a substantial portion of the variations in travel time.

2.4.3 Heterogeneity

Heterogeneity in HSR connectivity. We investigate the heterogeneous effects of HSR connections by introducing interactions between travel time and group variables that indicate specific HSR connection scenarios. The first group pertains to pairs directly connected to HSR with both regions having HSR stations. The second group consists of pairs where only one region has an HSR station. The third group encompasses pairs where neither region has an HSR station. Hence, three distinct β coefficients are estimated for each of these groups.

Heterogeneity in distance. The presence of an HSR station at both ends is not the sole source of heterogeneity in HSR connections. Various levels of distance and travel time also contribute to this heterogeneity. To account for this, we will incorporate different categories defined by specific distance and travel time thresholds. This approach will enable us to capture the nuanced effects of HSR connections across different geographical and commuting time contexts. We are also going to differentiate between different intensity of travel time reduction to gain additional insights into whether higher travel time reduction have stronger effects on patenting collaboration.

The Core and the Periphery. Investigating the differential effects of travel time reduction on co-patenting across core and periphery regions is essential for several reasons. Firstly, these categories represent distinct types of regions with varying levels of economic and innovative activities. Core regions, characterized by their urban nature, higher population size, and intensive patent activity, typically possess more developed research and innovation ecosystems. On the other hand, periphery regions, with their lower population size and limited patent activity, may have less mature innovation environments.

By examining how travel time reduction affects co-patenting in these different contexts, we can determine if travel time reduction has a more significant impact on co-patenting in pairs of core regions due to their already established innovative infrastructure. Alternatively, we can identify if periphery regions experience a more pronounced boost in co-patenting as they gain better access to core regions' innovation networks through improved connectivity. Core regions may leverage their existing research capacity to take advantage of reduced travel time for more frequent face-to-face interactions. In contrast, periphery regions might benefit from increased access to core regions' knowledge hubs, enabling them to tap into new sources of expertise and potentially elevate their own innovation capabilities.

In the regression analysis, we introduce interactions between travel time and three distinct groups: core-core, core-periphery, and periphery-periphery. These groups are defined based on Eurostat's NUTS regional urban classification, which categorizes regions according to their urbanization type. Eurostat's urban-rural typology classification classifies regions into three categories: (1) Predominantly Urban Regions, where at least 80% of the population resides in urban clusters with high population density - *15 regions*; (2) Intermediate Regions, with urban cluster populations between 50% and 80% of the total population, showing characteristics between predominantly urban and predominantly rural areas - *33 regions*; and (3) Predominantly Rural Regions, with at least 50% of the population living in rural grid cells characterized by

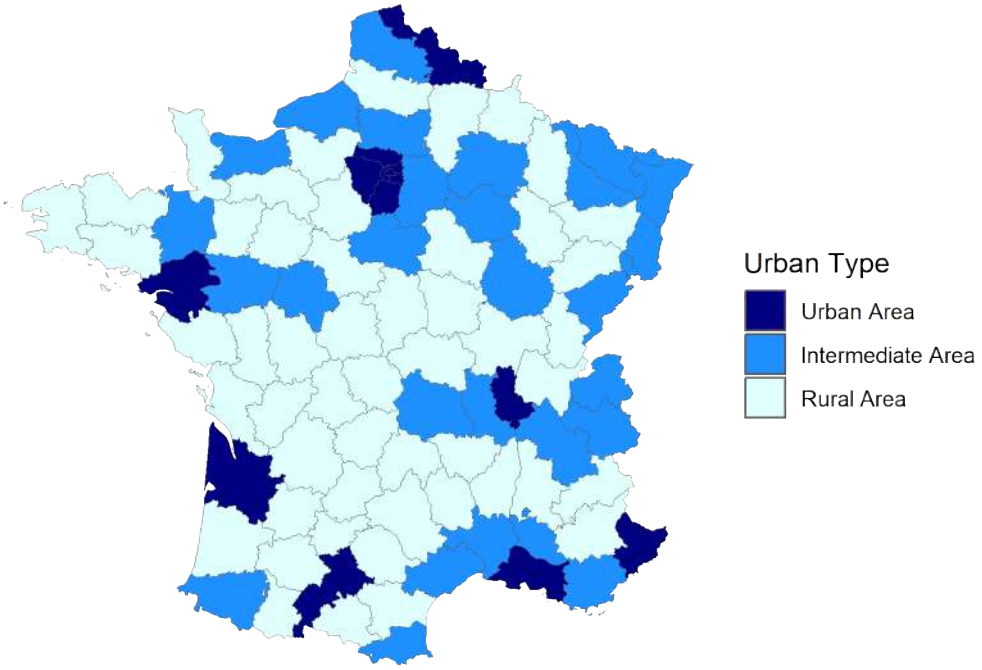


Figure 2.5: The Cores and the Periphery

low population density - *55 regions*. Figure 2.5 shows the map with the corresponding classification. The first two categories form the core regions, and the latter category represents the periphery. These distinctions provide insights into regional urbanization patterns, enabling us to assess how travel time reduction impacts collaboration in areas with varying levels of urbanization.

Technological sectors. The last heterogeneity that we explore is across technological fields, which are divided in 8 distinct groups according to the aggregated International Patent Classification: (1) *Human Necessities*, (2) *Performing, Operations and Transporting*, (3) *Chemistry and Metallurgy*, (4) *Textiles and Paper*, (5) *Fixed Constructions*, (6) *Mechanical Engineering, Lighting*, (7) *Physics*, and (8) *Electricity*. Different technological sectors may have varying degrees of sensitivity to travel time reduction. Some sectors might heavily rely on face-to-face interactions and knowledge exchange, making them more responsive to improved travel accessibility. On the other hand, sectors that are more digitized or less reliant on physical presence might show different patterns of collaboration. By analyzing the effects of travel time reduction in distinct sectors, we gain insights into which sectors benefit the most from enhanced connectivity.

To test our hypothesis regarding the significance of face-to-face interactions, we sought to identify technological sectors that potentially rely more on these in-person engagements. To accomplish this, we conducted an analysis of patents containing specific keywords in their abstracts, such as "experiment", "laboratory", "test" and "trial", both in English and French, to the singular and plural forms.[25] We calculate the proportion of patents featuring at least

[25] A similar methodology is mentionned in Bircan et al. (2022) to identify technologies relying on more face-to-face interactions. They find that chemistry patents involve the most experimentation, and electricity and fixed constructions patents the least.

Technology	IPC	Occurence (%)	Rank
Physics	G	3.18	1
Chemistry and Metallurgy	C	2.14	2
Textiles and Paper	D	1.44	3
Electricity	H	1.18	4
Human Necessities	A	1.00	5
Fixed Construction	E	0.87	6
Performing, Operations and Transporting	B	0.85	7
Mechanical Engineering, Lighting	F	0.67	8

The different technologicy classes correspond to the International Patent Classification (IPC). The table displays the frequency of keywords found in patent abstracts, signaling the requirement for face-to-face interactions. The identified keywords include "experiment," "laboratory," "test," and "trial" in both English and French, encompassing singular and plural forms.

Table 2.7: Ranking of face-to-face intensive technological fields

one of these keywords within each technological field. This computation is conducted using the iCrios Patstat database, which encompasses patents with inventors located in France from 1978 to 2016. The results, along with the corresponding rankings, are presented in Table 2.7.

The three first technological fields that count the higher amount of words related to the need of face-to-face interaction to conduct their work are the fields of Physics, Chemistry and Metalurgy, and Textiles and Paper.

2.4.4 The Nature, Quality and Mechanisms of Collaboration

First, we have explored the impact of travel time reduction on adjustments in the quantity of collaboration within region-pairs. Now, we delve into an investigation of both the extensive and intensive margins. Additionally, we examine how these reductions affect the quality of collaborations. Quality, in this context, is defined by various metrics, including the number of citations, the novelty of knowledge produced, measured by the number of claims within patents, and the scope of knowledge, often referred to as multi-disciplinarity, which is measured by the amount of different technological fields in these collaborations.

The nature of collaborations. We present similar regression estimations using various dependent variables: (1) The count of co-patents between regions i and j involving a newly established collaboration among inventors located in i and j. This represents collaborations that are initiated for the first time, relating to the extensive margin of collaboration patterns. (2) The count of co-patents between regions i and j involving pre-existing collaborations among inventors located in i and j. This category includes collaborations that have existed in the past and continue to yield co-patents, relateing to the intensive margin of collaboration. (3) The count of co-patents, weighted by the proportion of inventors within the pair contributing to each patent. This measure accounts for the collaborative effort within each patent. (4) The count of co-patents with only one owner or applicant, relating to intra-firm collaboration. (5) The count of co-patents with at least two different owners or applicants, indicating multi-party or inter-firm collaborations.

The quality of collaborations. Other dependant variables are used to assess the effect of travel time reduction on co-patents' quality: (1) the amount of citations per co-patent, (2) the amount of claims per co-patent, and (3) the amount of different technological fields per co-patent. The first measure assesses the quality of co-patents, using the number of citations as an indicator. A higher number of citations suggests that the patents have had a significant influence on subsequent inventions, highlighting their contribution to innovation and scientific advancement. The second measure serves as a proxy for the scope of the invention. It is based on the number of claims within a patent. A patent with a greater number of claims is indicative of higher novelty and complexity in the invention. Lastly, the third measure focuses on the diversity of different technological classes encompassed within a patent. This metric provides insights into the range of knowledge domains involved in a patent, reflecting the level of knowledge diversity.

Inventors' productivity within collaboration. Who is the HSR network connecting? We investigate the effect of travel time reduction on the amount of co-patents developed between different categories of inventors within region-pairs. We employ several different measures to investigate the patterns of co-patent development among inventors. First, we focus on the (1) count of co-patents developed by inventors who both exhibit productivity levels above their regional average, (2) where only one of the inventors surpasses the regional average, and (3) where both inventors fall below the average productivity level in their respective regions. We extend our analysis by considering co-patents involving inventors who fall within the top 10th percentile in terms of productivity within their respective regions. This investigation includes scenarios (4) where both inventors achieve this high level of productivity, as well as situations (5) where only one of them attains such a ranking.

Finally, we verify Catalini et al. (2020)'s prediction that improved connectivity to distant regions allows inventors to access a broader pool of potential collaborators, transcending geographical constraints to collaborate with inventors with higher productivity levels than those found within their local collaboration pool. To do so, we test the effect of travel time reduction on: (6) the count of co-patents between regions i and j, for which inventors in region j exhibit higher productivity than the average productivity in region i, and (7) the count of co-patents between regions i and j, for which inventors in region j demonstrating lower productivity than the average productivity in region i.

Knowledge similarity and complementarity. We investigate whether the HSR network helps connecting people with similar knowledge or complementary knowledge. In Column 1, the dependent variable reflects the count of co-patents between regions i and j, where inventors share a similar knowledge set. Meanwhile, in Column 2, the dependent variable captures the count of co-patents between regions i and j, where inventors possess complementary knowledge sets. We refer to *similar knowledge* if the inventors-pair-level technological similarity index between the two inventors collaborating is close to 1 (e.g., ≥ 0.8), and to *complementary knowledge* if the cosine similarity index value is below 0.8, implying that the inventors' patent portfolios cover distinct but related areas.

2.5 Results

This section presents the results.

2.5.1 Baseline

Naive Gravity

To begin, we provide evidence of the localized nature of co-patenting by initially estimating a naive gravity model, devoid of any fixed effects. Subsequently, we augment this analysis by incorporating region-year fixed effects, effectively accounting for all time-varying unobservable factors specific to each region within the pairs. This approach enables us to make meaningful comparisons between pairs and assess the impact of disparities in travel time on the intensity of co-patenting.

The results of this analysis are presented in Table 2.8. The initial piece of evidence is highlighted by the positive and statistically significant coefficient associated with the $same_{ij}$ variable. This variable pertains to intra-regional observations where $i = j$, indicating co-patenting activities within the same region. On average, this coefficient implies that there are approximately 17 times more co-patents developed locally, within a region's borders, in comparison to instances involving inter-regional collaborations.[26] The second corroborating evidence emerges from the positive and statistically significant coefficient linked to the $contiguous_{ij}$ variable, which applies to pairs of regions that share an administrative border. In essence, this coefficient indicates that there are more co-patents developed in collaboration with its contiguous neighbors than with inventors at longer distances.

Upon incorporating the distance variable into the regression analysis, we observe that the positive coefficients for the first two variables remain statistically significant. Furthermore, the coefficient linked to distance exhibits the expected negative sign. This implies that when comparing two pairs of regions, denoted as ij and ik, if region j is 10% closer to region i than region k is, we can anticipate that ij will have approximately 7% more co-patents than ik. The longer the distance between two regions, the less intense their collaborative activity. Substituting distance with travel time yields precisely the same coefficient.

Interestingly, when we introduce both distance and travel time into the analysis, the significance of distance diminishes, while the coefficient for travel time remains relatively stable at approximately -0.7. The time required to reach a potential collaborator appears to carry more significance than the mere geographical distance. This observation aligns with our hypothesis that reduced travel time can indeed foster increased collaborative activity. For instance, if regions j and k are equidistant from region i, but region j boasts a 10% shorter travel time, we can anticipate that ij will have approximately 7% more co-patents than ik.[27]

This naive gravity model furnishes evidence that collaboration exhibits a localized pattern, where the factor of travel time appears to play a more significant role than distance. This observation offers a suggestive indication that face-to-face interactions may hold substantial importance in fostering collaborative activities. Additionally, we learn that co-patenting increases with the number of inventors in the regions pairs while it is independent of patenting activity intensity, after controlling for distance or travel time.[28]

[26] $\exp^{2.83} = 16.95$

[27] This result holds when estimating an overdispersed Poisson model, where the variance of the dependant variable is higher than its average, which is the case for our data as shown by Table 2.3. The large amount of fixed effects in the structural gravity equation do not allow us to continue estimate the coefficients using the overdispersed Poisson model.

[28] The amount of inventors and patents is highly correlated. Within regions, a 1% increase in the amount of inventors is associated to a significant 0.97% increase in the amount of patent.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
(Intercept)	-6.52^{***}	-2.12^{***}	-2.07^{***}	-1.94^{***}				
	(0.26)	(0.51)	(0.55)	(0.55)				
same$_{ij}$	2.82^{***}	1.05^{***}	1.56^{***}	1.48^{***}	2.71^{***}	0.59^{**}	1.11^{***}	0.93^{***}
	(0.18)	(0.24)	(0.20)	(0.25)	(0.18)	(0.25)	(0.19)	(0.20)
contiguous$_{ij}$	2.31^{***}	1.16^{***}	1.22^{***}	1.18^{***}	2.37^{***}	1.00^{***}	1.02^{***}	0.94^{***}
	(0.09)	(0.10)	(0.12)	(0.13)	(0.09)	(0.10)	(0.14)	(0.11)
log(distance$_{ij}$)		-0.72^{***}		-0.06		-0.86^{***}		-0.14
		(0.06)		(0.13)		(0.07)		(0.14)
log(travel time$_{ijt}$)			-0.72^{***}	-0.69^{***}			-0.91^{***}	-0.81^{***}
			(0.07)	(0.12)			(0.08)	(0.15)
asinh(# patents$_{it}$ $\times$ # patents$_{jt}$)	0.33^{***}	0.05	0.03	0.02				
	(0.13)	(0.13)	(0.13)	(0.13)				
asinh(# inventors$_{it}$ $\times$ # inventors$_{jt}$)	0.35^{**}	0.59^{***}	0.56^{***}	0.57^{***}				
	(0.14)	(0.14)	(0.14)	(0.14)				
Num. obs.	256711	256711	256711	256711	211914	211914	211914	211914
Deviance	167330.19	157287.44	151341.21	151316.24	133711.81	122338.04	116815.72	116685.96
Log Likelihood	-117859.63	-112838.25	-109865.14	-109852.66	-101050.44	-95363.56	-92602.39	-92537.51
Pseudo R^2	0.77	0.78	0.78	0.78	0.78	0.79	0.80	0.80
Num. groups: dep_name_i-yr					2556	2556	2556	2556
Num. groups: dep_name_j-yr					2556	2556	2556	2556

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. Columns 5 to 8 include region-year fixed effects, γ_{it} and δ_{jt}, to control for all regional characteristics that varies over time and explain the amount of co-patents between two regions. Results show evidence that collaborations are mostly localized, with less occurence as distance and travel time increase.

Table 2.8: Naive Gravity

Structural Gravity

The structural gravity equation is estimated using the complete set of fixed effects, encompassing both the region-pair fixed effects and the two region-year fixed effects. Variables $same_{ij}$, $contiguous_{ij}$ and $distance_{ij}$ are accounted for within the region-pair fixed effects, and as a result, their effects are not directly estimated. The findings from this analysis are presented in Table 2.9. In the first column, we observe a negative relationship between travel time and co-patents. When travel time decreases by approximately 10%, which is close to the average reduction, we can expect an average increase of 3% in the number of co-patents within the affected pair.

	Model 1	Model 2	Model 3	Model 4	Model 5
log(travel time$_{ijt}$)	-0.29^{***}	0.24^{**}	-0.31^{**}	-0.41^{***}	-0.44^{**}
	(0.10)	(0.11)	(0.14)	(0.15)	(0.18)
Sample					
All pairs	Yes	No	No	No	No
Exclude pairs where $i = j$	No	Yes	Yes	Yes	Yes
Exclude pairs where i and j are contiguous	No	No	Yes	Yes	Yes
Exclude pairs where distance$_{ij} > 100$ km	No	No	No	Yes	Yes
Exclude pairs where distance$_{ij} > 200$ km	No	No	No	No	Yes
Num. obs.	122714	113840	98086	97124	75300
Num. groups: ij	4332	4242	3852	3830	3224
Num. groups: it	2556	2442	2322	2316	2158
Num. groups: jt	2556	2442	2322	2316	2158
Deviance	74436.32	66573.95	55505.71	54790.73	43002.26
Log Likelihood	−71412.70	−63765.29	−51649.39	−50543.43	−39576.77
Pseudo R^2	0.80	0.76	0.49	0.46	0.44

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. This specification allows for the identification of travel time reduction effect, within-pair. Results show evidence that as travel time decrease within pairs, the amount of co-patents increase. Columns 2 to 5 refine the regression sample by excluding certain pairs to assess the robustness of the travel time effect. This involves selecting specific pairs to be included in the pair fixed effects (intra-regional pairs and contiguous regions pairs) and increasing distance minimum thresholds.

Table 2.9: Structural Gravity

When we exclude pairs where $i = j$ from the sample used for estimation, the coefficient of travel time turns positive, as shown in column 2. However, upon further exclusion of pairs from contiguous regions, the coefficient reverts to negative, as seen in column 3. This result suggests that the reduction in travel time, which primarily occurs for inter-regional pairs, tends to increase the number of co-patents developed in comparison to intra-regional co-patents and all other pairs with no reduction in travel time.

However, when excluding intra-regional pairs, the comparison to contiguous pairs may distor the relationship because of possible agglomeration effects. Co-patenting tends to increase at a faster rate between contiguous regions than over long distances. As demonstrated in Table 2.4, the average growth rate in the number of co-patents between contiguous regions is approximately 244% from 1980 to 2010, while it's only about 31% for long-distance pairs. This trend holds true, with higher values, for pairs that have an HSR station at both ends or only at one end, although it remains lower than for contiguous pairs. We test this agglomera-

tion hypothesis in section **??**.

Furthermore, when we exclude pairs of regions within a 100km and 200km radius, we observe significant negative coefficients. Notably, these coefficients grow in magnitude as the distance threshold increases. Comparing pairs of region at high distance reveals that the greater the reduction in travel time between them, the more co-patents they tend to develop in collaboration. This effect is particularly pronounced when contrasted with pairs, similarly distant, that do not benefit from any reduction in travel time.

In summary, the results highlight a shift in collaboration patterns away from intra-regional collaboration towards pairs that have witnessed a reduction in travel time, attributed to the introduction of high-speed railways. This shift is particularly evident when comparing distant regions. However, when contrasting these pairs with contiguous ones, the effect becomes less pronounced, probably because of agglomeration patterns - inventors' locations gradually expanding in proximity to their workplace.

2.5.2 Communication Costs Proxies

Inventors' Interconnectivity and Technological Similarity

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
same$_{ij}$	1.06^{***}	0.59^{***}	0.61^{***}			
	(0.20)	(0.15)	(0.16)			
contig$_{ij}$	1.00^{***}	0.92^{***}	0.93^{***}			
	(0.14)	(0.12)	(0.12)			
log(travel time$_{ijt}$)	-0.89^{***}	-0.72^{***}	-0.73^{***}	-0.53^{***}	-0.23^{**}	-0.46^{***}
	(0.08)	(0.07)	(0.07)	(0.11)	(0.10)	(0.10)
asinh(bridges$_{ijt}$)	0.03		-0.01	0.12^{***}		0.13^{***}
	(0.02)		(0.01)	(0.01)		(0.01)
asinh(technosim$_{ijt}$)		3.20^{***}	3.21^{***}		0.62^{***}	0.80^{***}
		(0.22)	(0.22)		(0.14)	(0.15)
Num. obs.	211914	211914	211914	122714	122714	122714
Num. groups: ij				4332	4332	4332
Num. groups: it	2556	2556	2556	2556	2556	2556
Num. groups: jt	2556	2556	2556	2556	2556	2556
Deviance	116740.79	106931.19	106917.97	73736.69	74382.98	73649.89
Log Likelihood	-92564.93	-87660.13	-87653.52	-71062.88	-71386.02	-71019.48
Pseudo R^2	0.80	0.81	0.81	0.80	0.80	0.80

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. This table include additional communication costs proxies. Columns 1 to 6 include region-year fixed effects, γ_{it} and δ_{jt}, and colums 4 to 6 include pair fixed effects, ρ_{ij}. The three first specifications allow for between region-pairs evaluation, while the three last allow for within region-pairs evaluation. Results show evidence that elasticity to travel time is higher when accounting for the amount of bridges between two regions, which relates to the intensity of the inventors' network interconnectivity between regions i and j.

Table 2.10: Communication costs proxies

Table 2.10 presents the results, including additional proximity measures related to the collaboration network of inventors (*bridges*$_{ijt}$) and the similarity in technological field activity (*technosim*$_{ijt}$) across and within pairs. When considering a comparison across pairs, we find

that technological similarity exerts a stronger influence on co-patenting intensity than the presence of common collaborators among inventors, the latter of which does not exhibit any significant effect. However, when we incorporate pair fixed effects for a within-pair comparison, the influence of the latter variable becomes significantly positive. Thus, our expectations align with the observed results.

Furthermore, the coefficient for travel time remains significantly negative and becomes more pronounced with the inclusion of the variable $bridges_{ijt}$. This inflation in the travel time coefficient can be attributed to the correlation between these two variables. The higher magnitude of the travel time coefficient is in line with a negative correlation between the number of common collaborators and travel time, coupled with a positive correlation between the number of common collaborators and the quantity of co-patents. The negative correlation between travel time and the amount of common collaborators is especially significant for distant pairs as shown in Table 2.30. Improved transportation connections between two regions not only foster increased collaborations but also enhance the interconnectedness of those regions' inventors in terms of their collaboration network.

The present paper has successfully controlled for important covariates, mitigating the risk of underestimating the impact of high-speed railways due to omitted variable bias.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
log(travel time$_{ijt}$)	-0.26^{**}	-0.29^{***}	-0.29^{***}	-0.46^{***}	-0.29^{***}	-0.45^{***}
	(0.10)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)
waterways$_{ij}$ $\times t$	0.00	0.01^{***}				
	(0.00)	(0.00)				
asinh(bridges$_{ijt}$)		0.13^{***}		0.12^{***}		0.13^{***}
		(0.01)		(0.01)		(0.01)
asinh(technosim$_{ijt}$)		0.73^{***}		0.80^{***}		0.80^{***}
		(0.15)		(0.15)		(0.15)
$\mathbb{1}$(adsl$_{ijt} > 0$)			7.32^{*}	6.76^{*}		
			(3.91)	(3.72)		
$\mathbb{1}$(adsl$_{ijt} > 0$) $\times$ asinh(internet speed$_t$)			-1.44^{*}	-1.32^{*}		
			(0.75)	(0.71)		
adsl$_{ijt}$					0.44	1.74
					(2.93)	(2.83)
adsl$_{ijt}$ $\times$ asinh(internet speed$_t$)					-0.07	-0.32
					(0.46)	(0.44)
Num. obs.	122714	122714	122714	122714	122714	122714
Num. groups: ij	4332	4332	4332	4332	4332	4332
Num. groups: it	2556	2556	2556	2556	2556	2556
Num. groups: jt	2556	2556	2556	2556	2556	2556
Deviance	74434.47	73597.02	74427.97	73643.34	74436.26	73645.91
Log Likelihood	-71411.77	-70993.04	-71408.52	-71016.21	-71412.67	-71017.49
Pseudo R^2	0.80	0.80	0.80	0.80	0.80	0.80

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. This table include additional controls that may have decreased communication costs between region: regional coverage of internet $adsl_{ijt}$, and time trends interacted by waterways connection, waterways$_{ij} \times t$, where navigable waterways look alike the HSR network. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Results demonstrate the robustness of the elasticity to travel time. Additionally, the influence of internet access on collaboration is not observed to have a significant impact, as the effect of internet access is likely encompassed within the region-year fixed effects.

Table 2.11: Internet and waterways

New and Old Time Trends: Internet and Waterways

Controlling for time trends among long-distance pairs connected by waterways and considering internet connectivity does not diminish the significance of the impact of travel time on the number of co-patents developed within pairs of regions. The result holds when excluding regions directly desserved with an HSR station from the sample, as shown by table 2.31. Despite the increasing availability of telecommunication via the internet, one might have expected face-to-face interactions to lose significance. However, our findings indicate that they continue to play an important role, at least for the period of study.

Long-distance pairs connected by waterways are estimated to exhibit continuous average growth of approximately 1% per year.[29] This growth aligns with expectations, given the long history of interaction and relationship that those region-pairs have likely developed over time through their fluvial connection.

Interestingly, the effect of ADSL connectivity exhibits a positive association when internet speed is below 74 megabits per second but turns negative beyond this threshold. This unexpected result might be attributed to the nature of the internet access measure, which does not vary directly over time between pairs but rather within specific regions. This time change is already captured by the region-year fixed effects. Consequently, the negative coefficient may represent a non-linear effect, akin to a squared effect. It could imply that as internet speed increases, the rate of increase in co-patents developed decreases, suggesting decreasing marginal effect or diminishing returns of internet connection on collaboration.

2.5.3 Endogeneity

HSR connectivity

To examine whether the significant coefficient associated to the decrease in travel time is unaffected by endogeneity concerns arising from reverse causality, we narrow down the sample. First, we exclude pairs of regions where both have an HSR station, followed by excluding pairs where only one of the two regions has an HSR station. A similar approach has been undertaken by Kang et al. (2023). Results in table 2.12 show that even pairs without direct connection to HSR display a significant coefficient.

Lead and Lag effects

Table 2.13 presents the results of the lead and lag specification, which aims at exploring potential lead or lag effects associated with the implementation of HSR infrastructure. In this analysis, we regress the quantity of co-patents developed between pairs of regions on dummy variables that indicate the years before and after the first travel time reduction. We do not observe any lead or anticipation effect. This suggests that the HSR implementation do not depend on prior trends in collaboration. In other words, the decision to collaborate on patents does not seem to be influenced by past patterns of collaboration, providing evidence against endogeneity concerns related to the introduction of HSR infrastructure. It provides significant reassurance that the issue of HSR endogeneity arising from reverse causality is not a concern in our analysis.

[29] $100 \times (\exp(0.01) - 1) = 1.01$

	Model 1	Model 2	Model 3
log(traveltime)	-0.46^{***}	-0.44^{***}	-0.94^{***}
	(0.10)	(0.13)	(0.21)
asinh(technosim)	0.80^{***}	0.54^{***}	0.81^{***}
	(0.15)	(0.16)	(0.17)
asinh(bridges)	0.13^{***}	0.13^{***}	0.11^{***}
	(0.01)	(0.01)	(0.01)
Sample			
Both in HSR	Yes	No	No
One in HSR	Yes	Yes	No
Non in HSR	Yes	Yes	Yes
Num. obs.	122714	115470	73945
Num. groups: ij	4332	4148	2722
Num. groups: it	2556	2535	2080
Num. groups: jt	2556	2535	2080
Deviance	73649.89	68027.86	42196.11
Log Likelihood	−71019.48	−64821.51	−39384.92
Pseudo R^2	0.80	0.78	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Results demonstrate the robustness of the elasticity to travel time, even when excluding pairs that may pose endogeneity concerns—specifically, those where both regions have an HSR station and additionally those where only one region has an HSR station. Moreover, the elasticity of travel time exhibits a greater magnitude for pairs not directly connected by HSR.

Table 2.12: HSR endogeneity - selection of pairs according to the presence of HSR station

On the other hand, the analysis reveals that the effect of travel time reduction takes some time to materialize. There is no significant impact in the years immediately following the first travel time reduction, except for the sample of region-pair with no HSR station. After approximately four years and more, we observe substantial increases in co-patent activity. On average, co-patents are expected to increase by 15%, and for pairs without HSR stations in their respective regions, the increase is even more substantial at 25%. This delayed effect suggests that the full benefits of improved transportation connectivity may require a certain period to manifest in the form of collaborative innovation.

Ultimate Robustness: Inter-Regionalization Time Trends. To ensure that co-patenting is genuinely responsive to the travel time variable and not merely a result of time trends related to seeking external sources of knowledge for patent production, we introduce a border time trend into the regression. Additionally, to examine the increased propensity to collaborate with contiguous regions, in line with the agglomeration hypothesis, we incorporate a contiguity time trend into our analysis. Results are displayed in table 2.14.

The findings reveal that over time, regions are shifting their focus away from internal collaboration and increasingly engaging in inter-regional collaboration. Year by year, there is an average decrease in internal co-patents of approximately 3%-4%. This result holds statistical significance across all specifications, even when considering sample selection based on the presence of an HSR station in both regions involved in the pairs. Furthermore, there is an average annual increase in contiguous co-patents of approximately 0%-1%.

The inclusion of this time trends, especially the border time trend, decreases the magnitude

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$\mathbb{1}(t \leq t_0 - 3)$	-0.04	-0.06	0.02	-0.01	-0.04	0.02
	(0.03)	(0.04)	(0.06)	(0.04)	(0.04)	(0.06)
$\mathbb{1}(t = t_0 - 2)$	-0.05	-0.03	-0.00	-0.05	-0.05	-0.02
	(0.04)	(0.05)	(0.06)	(0.04)	(0.05)	(0.06)
$\mathbb{1}(t = t_0)$	-0.07	0.02	-0.06	-0.11^{**}	-0.02	-0.11^{*}
	(0.04)	(0.04)	(0.06)	(0.04)	(0.04)	(0.06)
$\mathbb{1}(t = t_0 + 1)$	0.04	0.06	0.13^{**}	0.01	0.03	0.10^{*}
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
$\mathbb{1}(t = t_0 + 2)$	-0.01	-0.02	-0.02	-0.03	-0.05	-0.04
	(0.03)	(0.04)	(0.05)	(0.03)	(0.04)	(0.05)
$\mathbb{1}(t = t_0 + 3)$	-0.02	0.04	0.08	-0.04	0.02	0.05
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
$\mathbb{1}(t \geq t_0 + 4)$	0.15^{***}	0.08^{**}	0.25^{***}	0.25^{***}	0.21^{***}	0.36^{***}
	(0.04)	(0.04)	(0.06)	(0.04)	(0.04)	(0.06)
Sample						
Both in HSR	Yes	No	No	Yes	No	No
One in HSR	Yes	Yes	No	Yes	Yes	No
Non in HSR	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	Yes	Yes	Yes
Num. obs.	122714	115470	73945	122714	115470	73945
Num. groups: ij	4332	4148	2722	4332	4148	2722
Num. groups: it	2556	2535	2080	2556	2535	2080
Num. groups: jt	2556	2535	2080	2556	2535	2080
Deviance	74387.90	68638.56	42412.98	73553.93	67965.49	42146.62
Log Likelihood	-71388.49	-65126.87	-39493.36	-70971.50	-64790.33	-39360.18
Pseudo R^2	0.80	0.78	0.68	0.80	0.78	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Co-patents are regressed on dummy variables that indicate the years before and after the first travel time reduction, identified as t_0. Hence $\mathbb{1}(t = t_0 + 1)$ equals 1, one year after the first travel time reduction, zero otherwise. $\mathbb{1}(t \geq t_0 + 4)$ equals 1, four years and more after the first travel time reduction, zero otherwise. Results show unsignificant lead effects - before travel time reduction - which reassures that we may not be concerned about endogeneity problems. Additionally, we find that the increase in co-patents takes four years and more to materialize.

Table 2.13: Lead and lag effects of travel time reduction

and the significance of the coefficient associated to travel time reduction. In particular, it becomes insignificant when using the full sample of pairs or when excluding pairs in the *Both in HSR* group only. However, the effect is statistically different from zero at the 10% level when using the sample of regions with no direct connection to the HSR. The results are driven by pairs of core regions, as showns in columns 4 to 6, where coefficient at statistically different from zero at the 5% level.

The inclusion of the time trend for each level of distance in table 2.35 show that, as time goes by, there are more co-patents developed for each distance value except within intra-regional borders. This effect erase the significance associated to travel time, except for *Core* regions not directly connected by HSR. However, it is possible that this last control is excessively robust and highly correlated with the widespread reduction in travel time across various periods and region pairs.

A potentially more revealing outcome may emerge with a less aggregated regional classi-

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
same$_{ij} \times t$	-0.03^{***}	-0.04^{***}	-0.04^{***}	-0.03^{***}	-0.04^{***}	-0.04^{***}
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
contiguous$_{ij} \times t$	-0.00	0.00	0.01^{*}	-0.00	0.00	0.01^{*}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
log(travel time$_{ijt}$)	-0.08	-0.20	-0.40^{*}			
	(0.11)	(0.13)	(0.21)			
log(travel time$_{ijt}$)×𝟙(Core-Core)				-0.13	-0.27^{**}	-0.53^{**}
				(0.11)	(0.14)	(0.24)
log(travel time$_{ijt}$)×𝟙(Core-Periphery)				0.40^{*}	0.16	-0.05
				(0.24)	(0.25)	(0.35)
log(travel time$_{ijt}$)×𝟙(Periphery-Periphery)				0.06	0.24	0.66
				(1.01)	(1.03)	(1.62)
asinh(bridges)	0.15^{***}	0.17^{***}	0.18^{***}	0.15^{***}	0.17^{***}	0.18^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
asinh(technosim)	0.27	-0.12	-0.04	0.29^{*}	-0.10	-0.03
	(0.16)	(0.15)	(0.18)	(0.16)	(0.15)	(0.18)
Sample						
Both in HSR	Yes	No	No	Yes	No	No
One in HSR	Yes	Yes	No	Yes	Yes	No
None in HSR	Yes	Yes	Yes	Yes	Yes	Yes
Num. obs.	122714	115470	73945	122714	115470	73945
Num. groups: ij	4332	4148	2722	4332	4148	2722
Num. groups: it	2556	2535	2080	2556	2535	2080
Num. groups: jt	2556	2535	2080	2556	2535	2080
Deviance	73013.99	67230.69	41622.46	73004.28	67225.78	41619.71
Log Likelihood	-70701.53	-64422.93	-39098.09	-70696.67	-64420.47	-39096.72
Pseudo R^2	0.80	0.78	0.68	0.80	0.78	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. In order to control for the general increase in inter-regional co-patenting practices, we control for time trends in pairs where $i = j$ and in contiguous pairs. Results find that as time goes by, there are less co-patents developed uniquely within a region's borders. This effect decrease the significance associated to travel time. However, significance remain for pairs involving *Core* regions, especially when they are not directly connected by HSR.

Table 2.14: Inter-regionalization time trends

fication. In a forthcoming version of this paper, we will employ commuting zones instead of NUTS3 regions. We anticipate that an analysis at a finer scale will emerge more differences in travel time reduction within same distance, allowing for better identification in the effect of travel time. We also anticipate that it could further amplify the anticipated influence of internet access, given that it would intensify the variations in its values among different pairs.

2.5.4 Heterogeneity

Access to High-Speed Railways

We examine how the impact of HSR connections varies across different groups depending on their connectivity to the network. The first group, *Both in HSR*, includes pairs directly connected to HSR, with both regions having HSR stations. The second group, *One in HSR*, comprises pairs where only one region has an HSR station. The third group, *None in HSR*, encompasses pairs where neither region has an HSR station.

Table 2.15 presents the results, where we estimate three distinct β coefficients for each of these groups. Therefore, it offers insights into the within-pair effects within each group. Col-

umn 2 includes the other communication costs proxies. Suprisingly, the reduction in travel time for pairs of regions directly connected to the HSR network does not appear to have a significant effect on the amount of co-patents developed. However, for the other two groups, the impact is statistically significant, particularly when accounting for the number of bridges (common collaborators). Interestingly, pairs of regions where neither region has an HSR station exhibit a higher magnitude in the elasticity coefficient compared to the other pairs of regions where one or both regions have a direct access to an HSR station.

We suggest different explanations for the insignificant coefficient of the former group. Regions that share a direct HSR connection might have already established collaboration patterns before the HSR network's introduction. In such cases, the reduction in travel time facilitated by HSR may not yield a significant additional benefit, resulting in a non-significant effect. It's also plausible that the impact of reduced travel time takes longer to manifest for pairs directly connected by HSR because they were already engaged in collaboration to some extent. Consequently, they may be less responsive to increased connectivity.

On the contrary, pairs without HSR stations at both ends but experiencing the advantages of the HSR network through reduced travel time along their route tend to display higher levels of responsiveness. The improvement in connectivity has the potential to transition them from a state where collaborating with other regions was cost-prohibitive, even though they recognized the value of collaborating with individuals from external knowledge sources, to a state where such collaboration becomes feasible and accessible.

Distance and Travel Time Thresholds

We examine the heterogeneous effects of travel time reduction based on the distance (expressed in kilometers) and the travel time in 2010. To do this, we establish various thresholds, and the results are presented in Table 2.16. Interestingly, for pairs with short distances (less than 100 kilometers) and short travel times in 2010 (below 1 hour), the effect of travel time reduction is not statistically significant. However, as the distance and travel time within reach increase, the effect of travel time on co-patenting intensity within pairs becomes more pronounced. For instance, a 10% reduction in travel time is associated with a 3.5% increase in the number of co-patents developed for pairs within a 100 to 200-kilometers reach. When considering the same decrease in travel time for pairs located over 200 kilometers apart, the associated increase in the number of co-patents is approximately 5%. This suggests that the impact of travel time reduction on co-patenting becomes more substantial as the geographical and temporal distances increase, which may be primarly due to very low levels of the amount of co-patents between distant pairs before the introduction of an HSR.

The Core and the Periphery

We investigate the differential impacts of travel time reduction on co-patenting across core and periphery regions. We introduce interaction terms between travel time and three distinct groups: core-core, core-periphery, and periphery-periphery. These groupings are defined using Eurostat's NUTS regional urban classification, as detailed in Section 2.4.3. See Figure 2.5 to visualize the classification of regions on a map. Dark to mid-blue refer to core regions, while light blue refer to the periphery. Table 2.17 presents the results of the econometric model.

Column 1 reveals a significant positive impact of travel time reduction on pairs of core regions, with an elasticity coefficient of -0.51. However, this reduction in travel time does

	Model 1	Model 2	Model 3
log(travel time$_{ijt}$)×𝟙(BothinHSR$_{ij}$)	-0.21	-0.22	-0.23^{*}
	(0.14)	(0.14)	(0.12)
log(travel time$_{ijt}$)×𝟙(OneinHSR$_{ij}$)	-0.36^{***}	-0.58^{***}	-0.56^{***}
	(0.14)	(0.14)	(0.14)
log(travel time$_{ijt}$)×𝟙(NoneinHSR$_{ij}$)	-0.32	-0.81^{***}	-0.75^{***}
	(0.22)	(0.22)	(0.22)
asinh(bridges$_{ijt}$)		0.13^{***}	
		(0.01)	
asinh(technosim$_{ijt}$)		0.81^{***}	
		(0.15)	
asinh(bridges$_{ijt}$)×𝟙(BothinHSR$_{ij}$)			0.11^{***}
			(0.02)
asinh(bridges$_{ijt}$)×𝟙(OneinHSR$_{ij}$)			0.16^{***}
			(0.01)
asinh(bridges$_{ijt}$)×𝟙(NoneinHSR$_{ij}$)			0.10^{***}
			(0.01)
asinh(technosim$_{ijt}$)×𝟙(BothinHSR$_{ij}$)			0.82^{**}
			(0.36)
asinh(technosim$_{ijt}$)×𝟙(OneinHSR$_{ij}$)			0.60^{***}
			(0.19)
asinh(technosim$_{ijt}$)×𝟙(NoneinHSR$_{ij}$)			0.88^{***}
			(0.17)
Num. obs.	122714	122714	122714
Num. groups: ij	4332	4332	4332
Num. groups: it	2556	2556	2556
Num. groups: jt	2556	2556	2556
Deviance	74433.94	73632.03	73558.07
Log Likelihood	-71411.51	-71010.55	-70973.57
Pseudo R^2	0.80	0.80	0.80

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Travel time is interacted by dummy variables which categorize region pairs into three groups based on their connection to high-speed railways: *Both in HSR* for pairs directly connected to HSR with both regions having stations, *One in HSR* for pairs with only one region having an HSR station, and *None in HSR* for pairs without HSR stations in either region. Results show higher magnitude in the coefficient associated to travel time for pairs with non-direct HSR connection.

Table 2.15: Heterogeneity in HSR connectivity

not exhibit a significant effect for pairs consisting of one core region and one periphery region, nor for pairs where both regions are classified as periphery. However, when we examine Column 2, we observe heterogeneous effects based on distance thresholds. We find that pairs of periphery regions experience substantial benefits from travel time reduction when the distance of separation is significant, exceeding 400 kilometers. The substantial magnitude of this latter coefficient can be attributed to the notably low levels of co-patenting observed between peripheral regions. Specifically, these levels were approximately 0.01 on average in 1980 and 0.1 in 2010. In contrast, during the same time frame, co-patenting increased from 0.5 to 3.5 for pairs of core regions on average, and from 0.02 to 0.3 for pairs consisting of one core and one

	Model 1	Model 2
asinh(bridges$_{ijt}$)	0.13^{***}	0.13^{***}
	(0.01)	(0.01)
asinh(technosim$_{ijt}$)	0.80^{***}	0.81^{***}
	(0.15)	(0.15)
log(travel time$_{ijt}$)×$\mathbb{1}$(distance$_{ij} \leq 100$)	-0.14	
	(0.22)	
log(travel time$_{ijt}$)×$\mathbb{1}$(100 < distance$_{ij} \leq 200$)	-0.35^{**}	
	(0.18)	
log(travel time$_{ijt}$)×$\mathbb{1}$(200 < distance$_{ij} \leq 400$)	-0.50^{***}	
	(0.15)	
log(travel time$_{ijt}$)×$\mathbb{1}$(distance$_{ij} > 400$)	-0.51^{***}	
	(0.15)	
log(travel time$_{ijt}$)×$\mathbb{1}$(travel time$_{ij,2010} \leq$ 1h)		-0.11
		(0.21)
log(travel time$_{ijt}$)×$\mathbb{1}$(1h < travel time$_{ij,2010} \leq$ 2h)		-0.34^{**}
		(0.14)
log(travel time$_{ijt}$)×$\mathbb{1}$(travel time$_{ij,2010} >$ 2h)		-0.61^{***}
		(0.13)
Num. obs.	122714	122714
Num. groups: ij	4332	4332
Num. groups: it	2556	2556
Num. groups: jt	2556	2556
Deviance	73645.35	73637.40
Log Likelihood	-71017.21	-71013.24
Pseudo R^2	0.80	0.80

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Travel time is interacted by distance and travel time thresholds. Results show higher magnitude in the coefficient associated to travel time for higher distance, and for higher travel time value as of 2010.

Table 2.16: Distance and travel time thresholds

periphery region.

The lack of significant benefits for core and periphery partnerships resulting from the high-speed rail network expansion is a concerning observation. This situation hampers the ability of periphery regions to tap into the innovative capacity of core regions. Several factors could explain this phenomenon. Firstly, it is worth noting that the national rail network may not be effectively integrate periphery regions, with trains operating at lower frequency compared to within and between core regions connections. This limited connectivity could hinder the potential benefits of reduced travel time. Additionally, several stations in the periphery have been shut down, further reducing accessibility. This closure initiative aimed to optimize rail services, enhance efficiency, and align with evolving transportation needs. Notably, between 1920 and 2020, around 40,000 km of rail lines were dismantled in France.

Furthermore, it is possible that improved HSR connections between core regions divert their collaborations away from periphery regions, as suggested by the positive but non-significant elasticity coefficient. This can further exacerbate the challenges faced by periphery regions in leveraging HSR for collaboration. It could be the reason why the results suggest an increased

	Model 1	Model 2
asinh(bridges$_{ijt}$)	0.12^{***}	0.13^{***}
	(0.01)	(0.01)
asinh(technosim$_{ijt}$)	0.81^{***}	0.81^{***}
	(0.15)	(0.15)
log(travel time$_{ijt}$)×𝟙(Core-Core)	-0.51^{***}	
	(0.11)	
log(travel time$_{ijt}$)×𝟙(Core-Periphery)	0.02	
	(0.24)	
log(travel time$_{ijt}$)×𝟙(Periphery-Periphery)	-0.32	
	(1.02)	
log(travel time$_{ijt}$)×𝟙(Core-Core) × 𝟙(distance$_{ij} \leq 200$)		-0.39^{**}
		(0.17)
log(travel time$_{ijt}$)×𝟙(Core-Core) × 𝟙($200 <$ distance$_{ij} \leq 400$)		-0.51^{***}
		(0.16)
log(travel time$_{ijt}$)×𝟙(Core-Core) × 𝟙(distance$_{ij} > 400$)		-0.61^{***}
		(0.16)
log(travel time$_{ijt}$)×𝟙(Core-Periphery) × 𝟙(distance$_{ij} \leq 200$)		0.65
		(0.42)
log(travel time$_{ijt}$)×𝟙(Core-Periphery) × 𝟙($200 <$ distance$_{ij} \leq 400$)		-0.47
		(0.31)
log(travel time$_{ijt}$)×𝟙(Core-Periphery) × 𝟙(distance$_{ij} > 400$)		0.22
		(0.38)
log(travel time$_{ijt}$)×𝟙(Periphery-Periphery) × 𝟙(distance$_{ij} \leq 200$)		1.93
		(1.91)
log(travel time$_{ijt}$)×𝟙(Periphery-Periphery) × 𝟙($200 <$ distance$_{ij} \leq 400$)		3.47^{**}
		(1.50)
log(travel time$_{ijt}$)×𝟙(Periphery-Periphery) × 𝟙(distance$_{ij} > 400$)		-4.13^{***}
		(1.32)
Num. obs.	122714	122714
Num. groups: ij	4332	4332
Num. groups: it	2556	2556
Num. groups: jt	2556	2556
Deviance	73640.25	73616.30
Log Likelihood	-71014.66	-71002.69
Pseudo R^2	0.80	0.80

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Travel time is interacted by dummy variables identifying groups of *Core-Core* regions, *Core-Periphery* regions, and *Periphery-Periphery* regions, as well as different distance thresholds. Results show that the significant effect of travel time only holds for pair of core regions, where the innovative activity is concentrated. Additionally, we find a higher magnitude in the coefficient associated to travel time for higher distance.

Table 2.17: The Core and The Periphery

partnership intensity between very distant periphery regions. They collaborate between each other rather than with cores. This may explain why the results indicate an uptick in partnership intensity among very distant periphery regions. In such cases, these periphery regions seem to prioritize collaboration. Lastly, the effects of HSR expansion on collaboration patterns may take time to fully materialize, as shown in Table 2.13. It's conceivable that the data analyzed do not capture the entire impact of the network's expansion within the studied time frame. Therefore, a longer-term perspective might be needed to assess the complete influence of HSR on regional collaborations.

Technological Fields

Figure 2.6 presents the results categorized by technological field. First figure employ region-pair fixed effects and incorporate region-year and technological field-year sets of fixed effects. This approach assumes that technological trends are homogenous across regions. for the second figure, we take a more granular approach by utilizing region-pair-technological fixed effects. This allows us to analyze the within pair-technology effect of travel time changes. Additionally, we incorporate region-technology-year fixed effects to account for varying trends associated with different technologies across diverse regions. Table 2.34 in the appendix shows the results with the selection of the sample according to the presence of an HSR station.

The results remain consistent across various technological fields, reaffirming the robustness of our previous findings. However, exceptions to this trend are observed in the fields of *Performing, Operations, and Transporting*, *Textiles and Paper*, *Fixed Construction* and *Physics*. Among pairs with no HSR station at both ends, *Textiles and Paper* and *Physics* fields find show a significant negative coefficient associated to travel time. This outcome aligns with our expectations, as these fields typically exhibit a lower reliance on face-to-face interactions as found in Table 2.7.

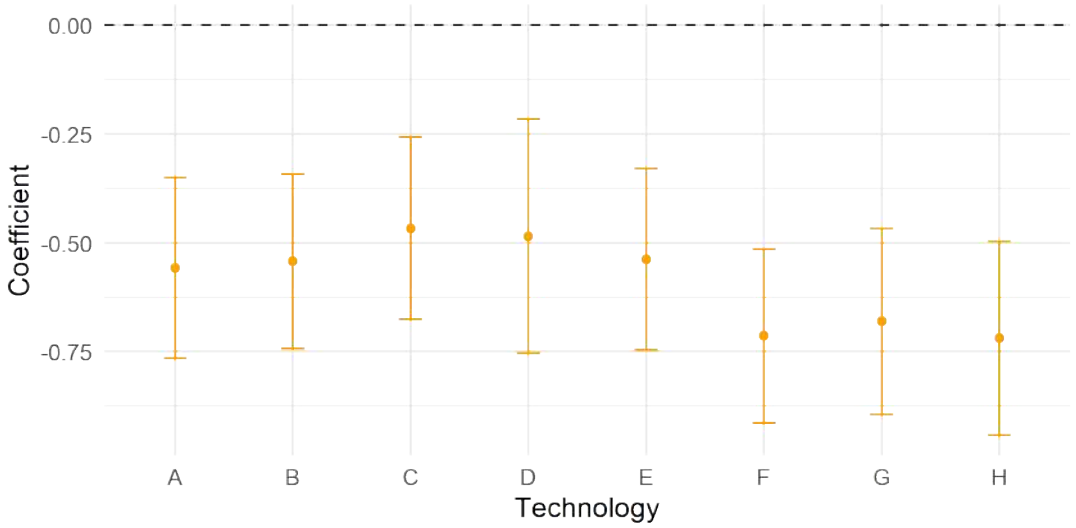
2.5.5 The Nature, Quality and Mechanisms of Collaboration

	Model 1 # copat$^{\text{new collab}}$	Model 2 # copat$^{\text{old collab}}$	Model 3 # copat$^{\text{share}}$	Model 4 # copat$^{\text{intra firm}}$	Model 5 # copat$^{\text{inter firm}}$
log(traveltime)	-0.47^{***}	-0.80^{***}	-0.32^{***}	-0.39^{***}	-0.70^{***}
	(0.10)	(0.24)	(0.10)	(0.12)	(0.19)
asinh(bridges)	0.10^{***}	0.38^{***}	0.10^{***}	0.14^{***}	0.02
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
asinh(technosim)	0.69^{***}	1.21^{***}	0.97^{***}	0.93^{***}	0.24
	(0.15)	(0.47)	(0.13)	(0.17)	(0.27)
Num. obs.	121742	32302	122714	108950	47794
Num. groups: ij	4328	1588	4332	3916	2462
Num. groups: it	2536	1705	2556	2496	1718
Num. groups: jt	2536	1705	2556	2496	1718
Deviance	70677.51	19560.59	45696.02	65083.74	23770.65
Log Likelihood	-67398.48	-18747.55	-49442.61	-61941.38	-21769.54
Pseudo R^2	0.78	0.72	0.87	0.80	0.40

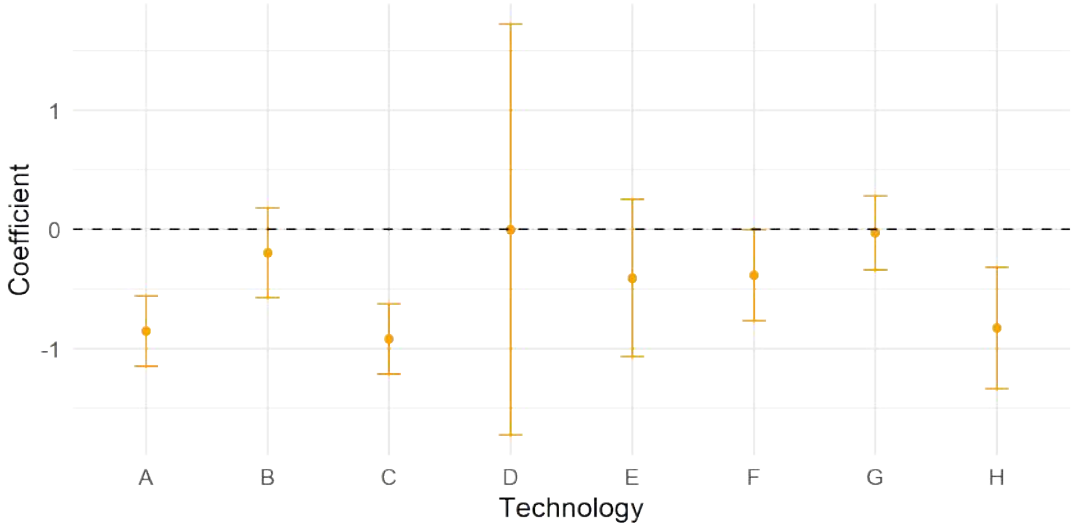
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variables used are: (1) the count of co-patents between regions i and j involving a newly established collaboration among inventors located in i and j; (2) the count of co-patents between regions i and j involving pre-existing collaborations among inventors located in i and j; (3) the count of co-patents, weighted by the proportion of inventors within the pair contributing to each patent; (4) the count of co-patents with only one owner or applicant, relating to intra-firm collaboration; and (5) the count of co-patents with at least two different owners or applicants, indicating multi-party or inter-firm collaborations. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}.

Table 2.18: Nature of Collaborations

The nature of collaborations. We conduct regression analyses with different dependent variables: (1) the amount of co-patents established from new collaboration between inventors, (2) co-patents developed with a continuation of pre-existing collaborations, (3) co-patents weighted by inventor contribution, (4) co-patents involving intra-firm collaboration, and (5)



(a) Fixed effects: FE_{ij}, FE_{it}, FE_{jt}, FE_{kt}



(b) Fixed effects: FE_{ijk}, FE_{ikt}, FE_{jkt}

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents in technological sector k developed between region i and j, identified by the location of inventors residence. International Patent Classification: (1) A: *Human Necessities*, (2) B: *Performing, Operations and Transporting*, (3) C: *Chemistry and Metallurgy*, (4) D: *Textiles and Paper*, (5) E: *Fixed Construction*, (6) F: *Mechanical Engineering, Lighting*, (7) G: *Physics*, and (8) H: *Electricity*. The identification includes pair, region-year and technology-year fixed effects in the first graph, as well as covariates. In the second graph, we include region-technology trio and technology-region-year fixed effects. Results show robust significant coefficient for the fields of *A: Human Necessities*, *C: Chemistry and Metallurgy*, *F: Mechanical Engineering, Lighting* and *H: Electricity*.

Figure 2.6: Travel time coefficient by technological field

inter-firm collaboration. The results in Table 2.18 consistently demonstrate the robustness of the travel time reduction effect across all specifications. Notably, there are fewer unique region-pairs in column 2 compared to column 1, indicating that the establishment of new collaborations extends to a broader geographical scope than the maintenance of pre-existing collaborations. This observation remains consistent when comparing column 5 to column 4, suggesting that there are fewer instances of inter-firm collaboration, or at the very least, inter-firm collaborations tend to be more localized geographically. This result may be attributed to the ease of communication and knowledge transfer within the confines of a firm, as discussed in the existing literature (Giroud et al., 2021).

	Model 1 # citations	Model 2 # claims	Model 3 # technology
log(travel time)	-0.31	-0.75^{***}	-0.56^{***}
	(0.30)	(0.14)	(0.11)
asinh(bridges)	0.11^{***}	0.11^{***}	0.04^{**}
	(0.02)	(0.02)	(0.02)
asinh(technosim)	-0.20	0.41^{**}	0.52^{***}
	(0.35)	(0.20)	(0.16)
Sample			
$t \leq 2005$	Yes	No	No
Num. obs.	13100	122714	122714
Num. groups: dep_name_i-dep_name_j	1807	4332	4332
Num. groups: dep_name_i-yr	1501	2556	2556
Num. groups: dep_name_j-yr	1501	2556	2556
Deviance	3311.24	118219.42	42236.67
Log Likelihood	-6057.81	-93613.14	-46271.44
Pseudo R^2	-0.39	0.39	0.19

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variables used are: (1) the average amount of citations per co-patent; (2) the average amount of claims per co-patent; (3) the average amount of technological fields per co-patent. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. The regression sample in column 1, using citations information in the dependent variable, spans from 1980 to 2005 due to the absence of 5-year forward citation data in the patents' application records beyond 2005. Results show that travel time reduction is associated to higher degree of novelty and multi-disciplinary research.

Table 2.19: Quality of Collaborations

The quality of collaborations. Table 2.19 presents the results. They indicate that the average number of citations per co-patent has not experienced a statistically significant increase as a result of travel time reduction, despite the observed negative coefficient of approximately -0.31. The average number of claims has risen for co-patents developed in response to the reduction in travel time, justifying the assumptions we made about the influence of external sources of knowledge on novelty. In particular, the improved connectivity seems to have facilitated the cross-pollination of ideas. Collaboration involving individuals from diverse backgrounds and regions can make innovative projects more novel and original as discussed and evidenced by Gallego et al. (2013). Finally, the reduction in travel time has heightened the multi-disciplinarity of co-patents, serving as additional evidence of the cross-pollination of ideas.

Inventors' productivity within collaboration. Results are displayed in table 2.20. Results

show that improved connectivity connects all inventor types, with an effect more pronounced for those with higher productivity — exceeding the local average and ranking in the top 10% within their region. Indeed, teams gathering two inventors surpassing their respective regional productivity average display coefficients with higher magnitude than teams involving only one or no inventors more productive than their regions' average. The magnitidue of the effect is even more pronounced when both are within the top 10% more productive inventors within their region.

	Top Average Productivity			*Top 10th Perc. Productivity*	
	both	*one*	*none*	*both*	*one*
	Model 1	Model 2	Model 3	Model 4	Model 5
log(travel time)	-1.16^{***}	-0.52^{**}	-0.26^{**}	-1.72^{***}	-1.24^{***}
	(0.24)	(0.22)	(0.13)	(0.64)	(0.26)
asinh(bridges)	0.36^{***}	0.18^{***}	0.00	0.76^{***}	0.32^{***}
	(0.02)	(0.01)	(0.01)	(0.05)	(0.02)
asinh(technosim)	1.00^{***}	-0.42	0.77^{***}	2.49^{*}	-0.12
	(0.31)	(0.27)	(0.17)	(1.32)	(0.37)
Num. obs.	34021	51501	73594	5623	34434
Num. groups: ij	1775	2552	3213	552	1885
Num. groups: it	1587	1757	2040	746	1601
Num. groups: jt	1587	1757	2040	746	1601
Deviance	20666.23	29234.75	38542.46	4077.41	19929.78
Log Likelihood	−18787.21	−27407.57	−36422.57	−4359.85	−18230.37
Pseudo R^2	0.67	0.67	0.72	0.57	0.63

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variables used are: (1) # copat$^{\text{both top avg}}$, (2) # copat$^{\text{one top avg}}$, (3) # copat$^{\text{no top avg}}$, (4) # copat$^{\text{both top 10th}}$, (5) # copat$^{\text{one top 10th}}$. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Results indicate that co-patents attributed to star inventors — those with productivity surpassing the local average and ranking in the top 10% within their region — are notably influenced by a reduction in travel time. In contrast, collaborations involving either both star inventors or no inventors with productivity above the local average exhibit less sensitivity to travel time reductions.

Table 2.20: Collaborations according to inventors' productivity

Finally, turning our attention to Table 2.21, our findings align with the model and empirical results put forth by Catalini et al. (2020). The results reveal that co-patents, which involve inventors with productivity levels exceeding the average productivity of their collaborator's pool, display a higher degree of sensitivity to travel time reduction compared to co-patents involving inventors whose productivity falls below the average of their collaborator's pool. In particular, the coefficient for the first is two times higher in magnitude than the second.

Knowledge similarity and complementarity. We investigate whether the HSR network helps connecting people with similar knowledge or complementary knowledge. The results indicate that HSR primarily facilitates connections between inventors who possess similar knowledge, as opposed to those with complementary knowledge. In particular, the coefficient in column (2) is statistically significant at the 10% level and its magnitude is approximately half that of column (1).

	# co-patents with inventor in region j has productivity	
	above region i's average	*below region i's average*
	Model 1	Model 2
log(traveltime)	-0.89^{***}	-0.43^{***}
	(0.21)	(0.14)
asinh(bridges)	0.29^{***}	0.06^{***}
	(0.02)	(0.01)
asinh(technosim)	0.76^{***}	0.74^{***}
	(0.27)	(0.15)
Num. obs.	46649	83015
Num. groups: ij	2256	3551
Num. groups: it	1837	2085
Num. groups: jt	1706	2094
Deviance	26422.55	45652.24
Log Likelihood	−24554.77	−44126.54
Pseudo R^2	0.70	0.78

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. In column 1, the dependent variable represents the count of co-patents between regions i and j, for which inventor(s) in region j exhibit higher productivity than the average productivity in region i. In column 2, the dependent variable captures the count of co-patents between regions i and j, for which inventor(s) in region j demonstrating lower productivity than the average productivity in region i. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. Results indicate that collaborations featuring inventors with productivity surpassing their collaborator's pool average are more responsive to travel time reduction, in contrast to collaborations involving inventors with productivity below the collaborator's pool average.

Table 2.21: Comparison to Catalini et al. (2020)

	Similar	*Complementary*
	Model 1	Model 2
log(traveltime)	-0.54^{***}	-0.25^{*}
	(0.11)	(0.14)
asinh(bridges)	0.14^{***}	0.15^{***}
	(0.01)	(0.01)
asinh(technosim)	1.06^{***}	−0.22
	(0.15)	(0.26)
Num. obs.	110939	70114
Num. groups: ij	3974	2898
Num. groups: it	2502	2135
Num. groups: jt	2502	2135
Deviance	63275.83	40033.01
Log Likelihood	−59209.61	−37739.55
Pseudo R^2	0.78	0.70

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. In column 1, the dependent variable represents the count of co-patents between regions i and j, for which inventors have a similar set of knowledge. In column 2, the dependent variable captures the count of co-patents between regions i and j, for which inventors have a complementary set of knowledge. Results show that improved connectivity primarily fosters connections among inventors with similar knowledge rather than those with complementary expertise.

Table 2.22: Similar VS Complementary Knowledge

2.6 Conclusion

This study explores the connection between transportation enhancements, particularly high-speed railways, and their influence on long-distance interactions manifested in inventors' collaborations on patents. The core premise is that the acceleration of travel times facilitated by high-speed railways promotes heightened mobility, subsequently fostering greater collaboration among inventors in interconnected regions. This hypothesis is based on the notion that such transportation systems create more opportunities and leaves more time for face-to-face teamwork on research projects.

Our paper, utilizing transportation network enhancements as quasi-natural experiments, sets apart from the literaure by yielding a unique travel time dataset for France. This dataset enables us to capture shifts in travel time subsequent to the introduction of high-speed railway lines. It allows us to provide external validity to the very recent studies that have investigated the impact of HSR on collaboration in China (Hanley et al., 2022; Li et al., 2022; Yao and Li, 2022; Kang et al., 2023).

Our econometric analysis, employing a gravity model with three-way fixed effects, revealed a significant increase in collaboration due to reduced travel times. This effect is especially pronounced for region-pairs not directly connected by high-speed rail, a result also pointed out by Kang et al. (2023). Robustness checks offer assurance that the reduction in travel time does not exhibit a lead effect. This finding alleviates concerns about potential endogeneity issues stemming from the construction of high-speed rail between regions with increasing collaboration trends to further boost this trend. Thus, it suggests that HSR development was not strategically aimed at stimulating collaboration.

We incorporated a crucial control variable, inspired by network literature Bergé (2015): the amount of bridges, which represent the number of common collaborators between regions based on their respective inventors' network. This measure is correlated with travel time improvements. As travel time decrease, the interconnectedness in the network of inventors improves. This measure substantially mitigates omitted variable bias, preventing an upward bias estimate (a downward bias in the coefficient's magnitude), thus enhancing the accuracy of the estimators.

Additionally, we account for internet connectivity, which appears to have an insignificant impact on co-patenting intensity, likely because region-year fixed effects inherently control for such variables. Furthermore, we include controls for time trends in region-pairs connected by waterways, given the similarities between navigable waterways and high-speed rail networks, as evidenced in the appendix. Notably, the effect of changes in travel time remains robust to these specifications.

Our heterogeneity analysis investigates the impact of travel time reduction on different distance thresholds, different groups of regions, and different technology sectors. We find that travel time reduction had more effect on long-distance pairs, at more than 200 kilometers, and for pairs of core regions, characterized by higher innovative activity compared to the periphery.

We not only find that the elasticity coefficient of travel time increases as regions are more distant, but also that the predicted increase in co-patents becomes a more significant factor in explaining the observed co-patent growth. When considering pairs of core regions and different distance thresholds (below 200 kilometers, between 200 and 400 kilometers, and above 400 kilometers), the predicted increase in co-patents following reduced travel time accounts for

0.5%, 9%, and 20.7% of the observed co-patent growth in core regions, within these respective distance ranges.

In the section where we intend to uncover the mechanisms behind this expansion of collaboration across regions, we discover that it contributed to a more diverse pool of ideas and expertise, ultimately enriching the scope and multidisciplinary nature of inventions, although it did not necessarily result in more citations. We also demonstrate that high-speed railways not only strengthened existing collaborations, but also facilitated new ones.

Moreover, we considered the role of star inventors, which demonstrate higher productivity levels, as defined by the amount of forward citations in the patents they have contributed to (Akcigit et al., 2018). We find that high-speed railways allowed to connect all types of inventors, but star inventors exhibited greater sensitivity to travel time reductions. This finding implies that inventors are willing to bear the costs of distance for enhanced innovation potential inherent in collaborating with highly productive inventors, as modeled by Catalini et al. (2020).

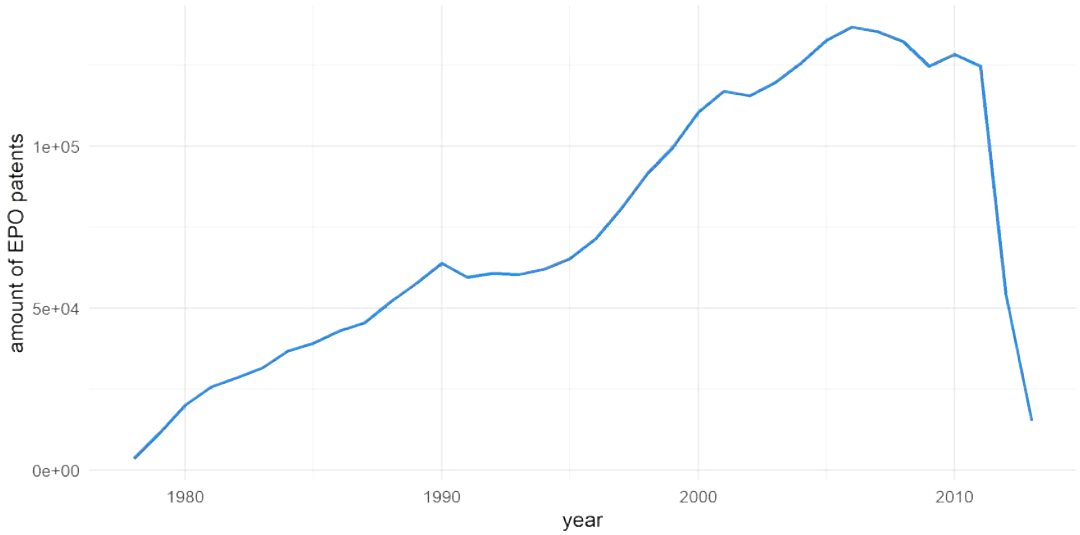
In line with Catalini et al. (2020), we delve deeper into the analysis by examining the impact of travel time on the number of co-patents between regions i and j, when inventors in region j collaborating with those in region i exhibit higher productivity than the average productivity of inventors in region i. Our findings affirm that inventors are indeed more responsive to travel time reductions when seeking collaborators with superior productivity compared to their local pool of inventors.

To sum up, our findings highlight the transformative potential of high-speed railways as catalysts for innovation. They serve as more than just transportation networks; they act as enablers of innovation, connecting and nurturing innovation hubs. Policymakers should take note of these results and consider investing strategically in transportation infrastructure to efficiently link innovation hubs, unlocking fresh opportunities for thriving innovation ecosystems.

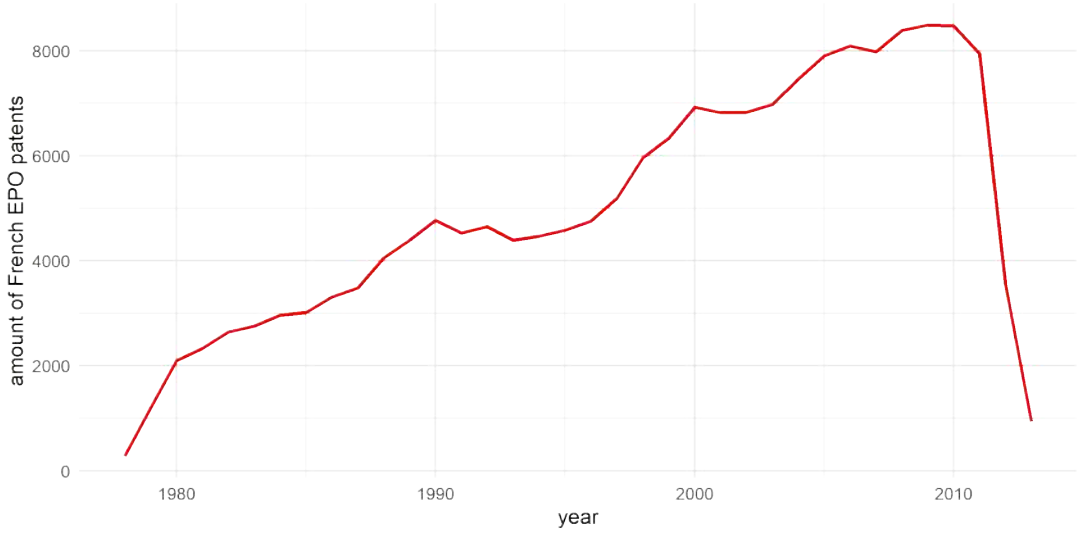
Nevertheless, our study also raises concerns about the potential divide between core and peripheral regions, as the benefits of the HSR network appear to be more pronounced in the former. Policymakers should consider measures to ensure that peripheral areas can also tap into the innovation potential offered by enhanced connectivity.

Appendix to chapter 2

2.A Appendix: Patent Data



(a) All patents



(b) French patents

Figure 2.7: Yearly count of inventions submitted at EPO

	Min	1st qu.	Median	Mean	3rd qu.	Max
All	0.15	1	1	0.97	1	1
Conditional on non-complete information	0.15	0.67	0.67	0.69	0.75	0.94

Table 2.23: Proportion of inventors' location information within patents teams

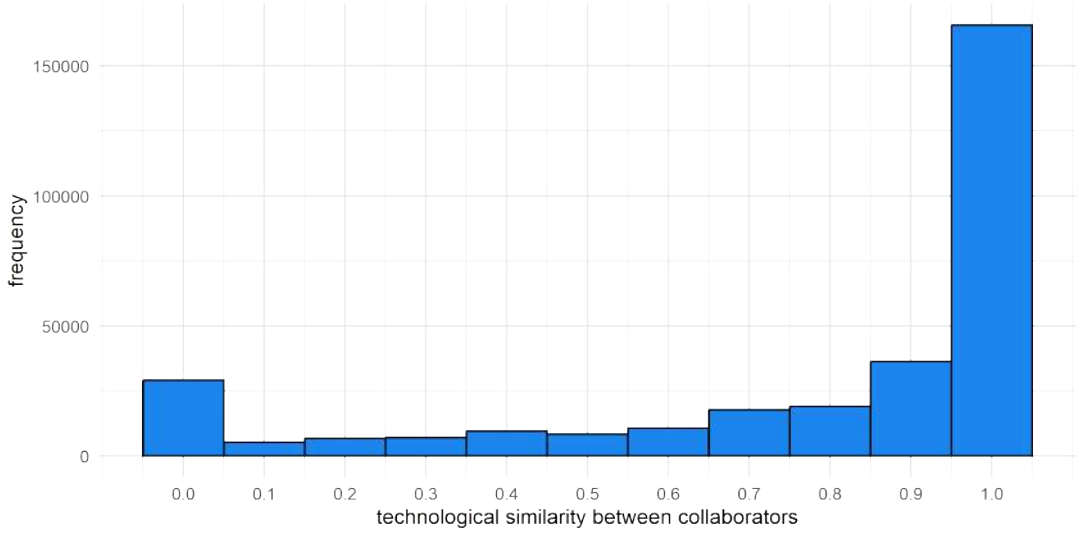


Figure 2.8: Technologcial similarity between collaborators

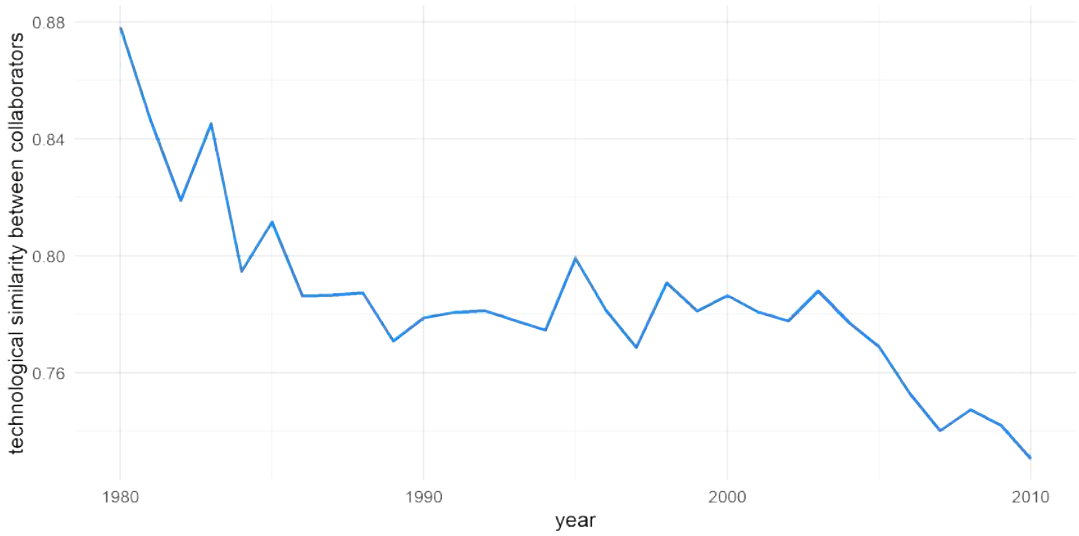


Figure 2.9: Time trends of average technologcial similarity between collaborators

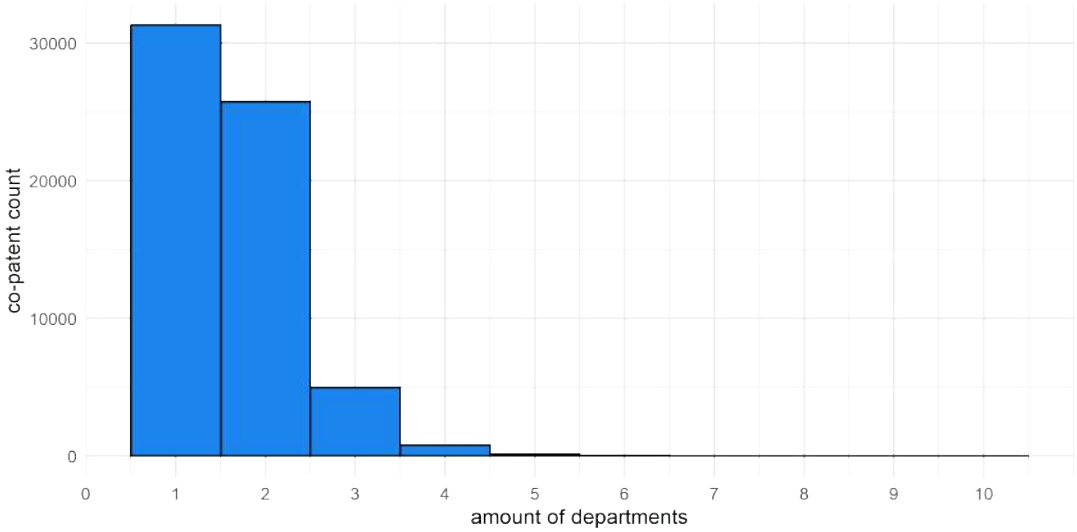


Figure 2.10: # NUTS3 regions involved within co-patents

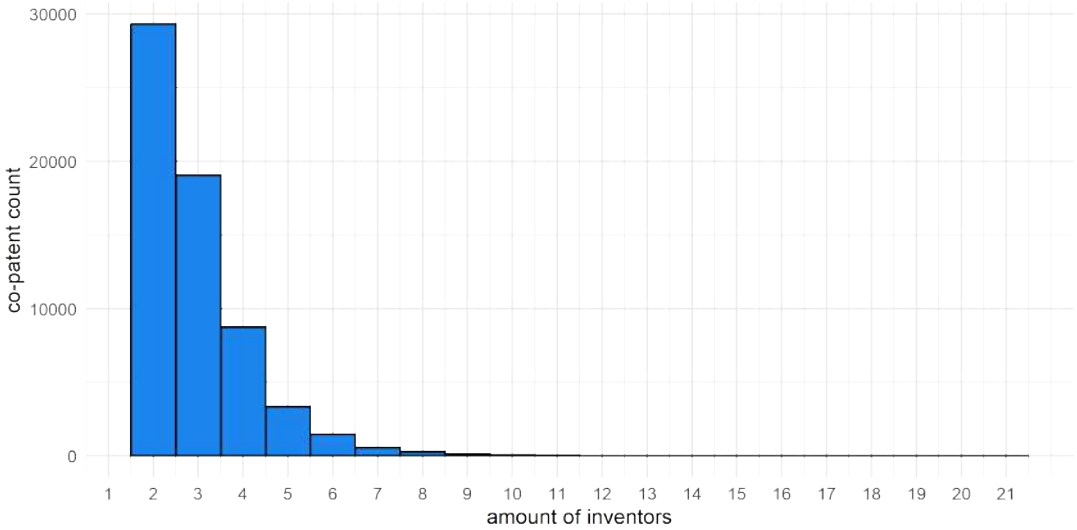


Figure 2.11: # inventors involved within co-patents

	Model 1	Model 2	Model 3
# inventors	-0.13^{***}		-0.13^{***}
	(0.00)		(0.00)
$\mathbb{1}$(inter-regional)		-0.01^{***}	0.06^{***}
		(0.00)	(0.00)
Num. obs.	64398	64398	64398
Num. groups: year	31	31	31
Adj. R^2 (full model)	0.17	0.00	0.17
Adj. R^2 (proj model)	0.17	0.00	0.17

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.24: Likelihood for a patent to have complete infomation on inventors' location

	Min	1st qu.	Median	Mean	3rd qu.	Max	N
Full count	1	1	1	2.23	2	156	77,345
Count weighted by participation	0.05	0.33	0.5	1.03	0.83	249.93	77,345
Average productivity	0	0	0	0.68	1	42.78	59,560
Maximum productivity	0	0	0	0.84	1	81	59,560

Table 2.25: Patent count per inventor

2.B Appendix: Tables

	Min	1st qu.	Median	Mean	3rd qu.	Max	N
Full count	1	1	1	2.27	2	150	56,022
Count weighted by participation	0.07	0.33	0.5	0.82	0.83	65.42	56,022
Average productivity	0	0	0	0.80	1	44.70	41,557
Maximum productivity	0	0	0	0.97	1	81	41,557

Table 2.26: Patent count per leader

Min	1st qu.	Median	Mean	3rd qu.	Max	N
0	0.67	0.96	0.77	1	1	316,296

Table 2.27: Technological similarity between inventors

	Min	1st qu.	Median	Mean	3rd qu.	Max
All	2	2	3	3.38	4	21
Intra-regional	2	2	3	2.92	3	12
Inter-regional	2	2	3	3.76	5	21

Table 2.28: Summary statistics on the # of inventors involved in collaboration within co-patents

2.C Appendix: Waterways

2.C.1 Introduction

In France, the public domain of rivers encompasses around 18,000 km of water routes, out of which around 8,500 km are navigable. These navigable routes include both natural waterways like rivers and streams, as well as artificial waterways such as canals, suitable for navigation by ship. They serve various purposes, including transportation of goods, passengers, recreational activities, tourism, and irrigation. For centuries, France has been engaged in constructing canals and improving rivers to establish navigable waterways. Here are some of the key types of navigable waterways in France:

1. Rivers: they have historically served as important trade routes. Rivers like la Seine (777 km), La Loire (1,006 km), la Garonne (529 km), le Rhône (814 km) et le Rhin (1,233km) have been crucial for transporting goods and connecting different regions. These rivers have been subject to various improvements, such as dam construction, and the creation of navigation channels, to enhance their navigability.

2. Canals: they are artificial waterways constructed to connect different regions or rivers. Canals provide an alternative transportation route, bypassing natural obstacles and facilitating the movement of goods. Famous canals in France include the Canal du Midi, which connects Toulouse to the Mediterranean Sea (240 km) since the 17th century, the Canal de Bourgogne, linking the Seine basin to the Rhône basin (242 km) since 1832, and the Canal du Rhône au Rhin, connecting the seaports of northern Europe with those of the Mediterranean by creating a Rotterdam-Marseille river link (375 km) since 1833.

3. Seaways and Coastal Waters: France has an 5,500 km of coastline along the North Sea, English Channel, Atlantic Ocean, and the Mediterranean Sea. These coastal waters provide important navigable routes for maritime transportation, connecting France to other countries and facilitating international trade. Coastal waters can also play a role in domestic trade. In some cases, navigable rivers flow into coastal waters, allowing continuity of transport from/to inland regions served by waterways. For instance, la Seine and Garonne rivers can connect inland ports in Paris and Bordeaux through coastal routes.

Centuries ago, France embarked on the development of navigable waterways, including the renowned Canal du Midi, which connected the Atlantic Ocean to the Mediterranean Sea.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
log(travel time$_{ijt}$)	-0.46^{***}	0.04	-0.23	-0.23^{**}	0.18	-0.31^{**}
	(0.10)	(0.11)	(0.14)	(0.10)	(0.11)	(0.14)
asinh(bridges$_{ijt}$)	0.13^{***}	0.21^{***}	0.25^{***}			
	(0.01)	(0.01)	(0.01)			
asinh(technosim$_{ijt}$)	0.80^{***}	-0.06	-0.11			
	(0.15)	(0.16)	(0.19)			
asinh(bridges$_{ij,t-1}$)				-0.00	0.05^{***}	0.02^{**}
				(0.01)	(0.01)	(0.01)
asinh(technosim$_{ij,t-1}$)				0.47^{***}	-0.36^{**}	-0.28
				(0.13)	(0.15)	(0.19)
Sample						
All pairs	Yes	No	No	Yes	No	No
Exclude pairs where $i = j$	No	Yes	Yes	No	Yes	Yes
Exclude pairs where i and j are contiguous	No	No	Yes	No	No	Yes
Num. obs.	122714	113840	98086	122714	113840	98086
Num. groups: dep_name_i-dep_name_j	4332	4242	3852	4332	4242	3852
Num. groups: dep_name_i-yr	2556	2442	2322	2556	2442	2322
Num. groups: dep_name_j-yr	2556	2442	2322	2556	2442	2322
Deviance	73649.89	65371.11	54602.55	74401.33	66493.30	55495.45
Log Likelihood	-71019.48	-63163.87	-51197.82	-71395.20	-63724.96	-51644.27
Pseudo R^2	0.80	0.76	0.49	0.80	0.76	0.49

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.29: Communication Costs Proxies 2

	Bridges			Technological Similarity		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
log(travel time)	-0.12	-1.10	-2.25^{**}	-0.09^{**}	0.04	0.03
	(0.28)	(0.71)	(0.94)	(0.04)	(0.03)	(0.03)
Sample						
All pairs	Yes	No	No	Yes	No	No
Exclude pairs where $i = j$	No	Yes	Yes	No	Yes	Yes
Exclude pairs where i and j are contiguous	No	No	Yes	No	No	Yes
Num. obs.	14695	11138	7698	256711	243254	229620
Num. groups: dep_name_i dep_name_j	980	908	750	8281	8190	7732
Num. groups: dep_name_i yr	1151	959	780	2821	2757	2757
Num. groups: dep_name_j yr	1151	959	780	2821	2757	2757
Deviance	20608.09	12668.98	8512.71	9179.74	7762.65	7312.22
Log Likelihood	-16759.63	-11110.80	-7385.18	-155711.75	-152182.20	-143065.55
Pseudo R^2	0.87	0.78	0.48	0.02	-0.00	-0.01

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.30: Effect of travel time reduction on inventors' interactions and specialization

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
log(travel time)	-0.58^{***}	-0.79^{***}	-0.59^{***}	-0.94^{***}	-0.60^{***}	-0.95^{***}
	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)
waterways$_{ij} \times t$	0.00	0.01^{**}				
	(0.00)	(0.00)				
asinh(bridges)		0.11^{***}		0.11^{***}		0.11^{***}
		(0.01)		(0.01)		(0.01)
asinh(technosim)		0.78^{***}		0.81^{***}		0.81^{***}
		(0.17)		(0.17)		(0.17)
adslpop_ij > 0			10.55	10.45		
			(8.15)	(8.04)		
adslpop_ij > 0):asinh(internet.speed)			-2.03	-2.01		
			(1.53)	(1.51)		
adslpop_ij					-10.87^{*}	-9.05
					(5.69)	(5.79)
adslpop_ij:asinh(internet.speed)					1.61^{*}	1.24
					(0.85)	(0.87)
Num. obs.	73945	73945	73945	73945	73945	73945
Num. groups: dep_name_i-dep_name_j	2722	2722	2722	2722	2722	2722
Num. groups: dep_name_i-yr	2080	2080	2080	2080	2080	2080
Num. groups: dep_name_j-yr	2080	2080	2080	2080	2080	2080
Deviance	42450.26	42182.50	42445.53	42191.38	42444.64	42188.87
Log Likelihood	-39512.00	-39378.11	-39509.63	-39382.55	-39509.19	-39381.30
Pseudo R^2	0.68	0.68	0.68	0.68	0.68	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.31: Internet and Waterways, sample with no HSR station in i and j

	Model 1	Model 2	Model 3
log(travel time)	-0.14	-0.35^{**}	-0.56^{***}
	(0.15)	(0.14)	(0.22)
asinh(bridges)	0.11^{***}	0.12^{***}	0.11^{***}
	(0.01)	(0.01)	(0.01)
asinh(technosim)	1.13^{***}	0.87^{***}	0.96^{***}
	(0.18)	(0.16)	(0.15)
waterways$_{ij} \times t$	0.00	0.01^{***}	0.00^{*}
	(0.00)	(0.00)	(0.00)
adslpop_ij > 0	4.33	7.29^{*}	12.95^{***}
	(4.68)	(4.03)	(4.51)
adslpop_ij > 0):asinh(internet.speed	-0.86	-1.40^{*}	-2.46^{***}
	(0.89)	(0.77)	(0.86)
Sample			
Both in HSR	Yes	No	No
One in HSR	No	Yes	No
None in HSR	No	No	Yes
$\mathbb{1}(\Delta^{1980,2010}\text{TravelTime} = 0)$	Yes	Yes	Yes
Num. obs.	37842	66665	84087
Num. groups: dep_name_i-dep_name_j	1416	2411	3039
Num. groups: dep_name_i-yr	2454	2503	2534
Num. groups: dep_name_j-yr	2455	2504	2536
Deviance	25553.99	43344.80	49117.96
Log Likelihood	-30234.84	-45460.67	-48184.89
Pseudo R^2	0.87	0.84	0.84

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.32: HSR connectivity heterogeneity, internet and waterways access

	Model 1	Model 2	Model 3
$\mathbb{1}(t \leq t_0 - 3)$	-0.02	-0.05	0.05
	(0.04)	(0.04)	(0.06)
$\mathbb{1}(t = t_0 - 2)$	-0.04	-0.04	-0.00
	(0.04)	(0.05)	(0.06)
$\mathbb{1}(t = t_0)$	-0.08^{*}	-0.00	-0.07
	(0.04)	(0.04)	(0.06)
$\mathbb{1}(t = t_0 + 1)$	0.02	0.04	0.13^{**}
	(0.04)	(0.04)	(0.05)
$\mathbb{1}(t = t_0 + 2)$	0.00	-0.03	-0.01
	(0.03)	(0.04)	(0.05)
$\mathbb{1}(t = t_0 + 3)$	-0.01	0.03	0.08
	(0.04)	(0.04)	(0.05)
$\mathbb{1}(t = t_0 + 4)$	0.14^{***}	0.13^{***}	0.25^{***}
	(0.04)	(0.04)	(0.07)
asinh(bridges)	0.15^{***}	0.17^{***}	0.18^{***}
	(0.01)	(0.01)	(0.01)
asinh(technosim)	0.24	-0.15	-0.08
	(0.16)	(0.15)	(0.18)
$\text{same}_{ij} \times t$	-0.02^{***}	-0.04^{***}	-0.04^{***}
	(0.01)	(0.00)	(0.01)
$\text{contiguous}_{ij} \times t$	0.00	0.00^{**}	0.01^{**}
	(0.00)	(0.00)	(0.00)
Sample			
Both in HSR	Yes	No	No
One in HSR	Yes	Yes	No
Non in HSR	Yes	Yes	Yes
Num. obs.	122714	115470	73945
Num. groups: dep_name_i-dep_name_j	4332	4148	2722
Num. groups: dep_name_i-yr	2556	2535	2080
Num. groups: dep_name_j-yr	2556	2535	2080
Deviance	72968.10	67193.56	41586.03
Log Likelihood	-70678.59	-64404.37	-39079.88
Pseudo R^2	0.80	0.78	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.33: Lead and lag effects of travel time reduction

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
log(travel time)×$\mathbb{1}(k=1)$	-0.56^{***}	-0.54^{***}	-0.85^{***}	-0.85^{***}	-0.69^{***}	-0.90^{**}
	(0.11)	(0.14)	(0.22)	(0.15)	(0.21)	(0.38)
log(travel time)×$\mathbb{1}(k=2)$	-0.54^{***}	-0.56^{***}	-0.99^{***}	-0.20	0.30	0.39
	(0.10)	(0.14)	(0.22)	(0.19)	(0.23)	(0.45)
log(travel time)×$\mathbb{1}(k=3)$	-0.47^{***}	-0.47^{***}	-0.74^{***}	-0.92^{***}	-0.96^{***}	-1.10^{***}
	(0.11)	(0.14)	(0.22)	(0.15)	(0.19)	(0.31)
log(travel time)×$\mathbb{1}(k=4)$	-0.48^{***}	-0.45^{***}	-0.92^{***}	-0.00	-2.38^{*}	-4.46^{**}
	(0.14)	(0.15)	(0.24)	(0.88)	(1.22)	(1.79)
log(travel time)×$\mathbb{1}(k=5)$	-0.54^{***}	-0.54^{***}	-1.00^{***}	-0.41	-0.84^{*}	-1.28
	(0.11)	(0.14)	(0.22)	(0.34)	(0.48)	(0.88)
log(travel time)×$\mathbb{1}(k=6)$	-0.71^{***}	-0.73^{***}	-1.16^{***}	-0.39^{**}	-0.51^{*}	-1.87^{***}
	(0.10)	(0.14)	(0.22)	(0.20)	(0.28)	(0.63)
log(traveltime)×$\mathbb{1}(k=7)$	-0.68^{***}	-0.66^{***}	-1.08^{***}	-0.03	-0.43^{**}	-0.67^{*}
	(0.11)	(0.14)	(0.24)	(0.16)	(0.22)	(0.36)
log(travel time)×$\mathbb{1}(k=8)$	-0.72^{***}	-0.69^{***}	-1.15^{***}	-0.83^{***}	-1.03^{***}	-0.99^{**}
	(0.11)	(0.14)	(0.25)	(0.26)	(0.33)	(0.47)
asinh(bridges)	0.13^{***}	0.13^{***}	0.10^{***}	0.13^{***}	0.12^{***}	0.10^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
asinh(technosim)	0.74^{***}	0.48^{***}	0.87^{***}	0.40^{**}	0.08	0.33^{*}
	(0.16)	(0.17)	(0.18)	(0.17)	(0.18)	(0.19)
Sample						
Both in HSR	Yes	No	No	Yes	No	No
One in HSR	Yes	Yes	No	Yes	Yes	No
None in HSR	Yes	Yes	Yes	Yes	Yes	Yes
Num. obs.	929160	878656	574048	291879	258623	151895
Num. groups: dep_name_i-dep_name_j	4112	3955	2647			
Num. groups: dep_name_i-yr	2493	2473	2049			
Num. groups: dep_name_j-yr	2493	2473	2049			
Num. groups: ipc-yr	248	248	248			
Num. groups: dep_name_i-dep_name_j-ipc				14285	13485	8519
Num. groups: dep_name_i-ipc-yr				12462	11938	9212
Num. groups: dep_name_j-ipc-yr				12462	11938	9212
Deviance	269620.62	244066.17	150224.22	158323.41	140421.64	82317.01
Log Likelihood	-206718.16	-185604.61	-111459.90	-151069.56	-133782.34	-77506.30
Pseudo R^2	0.70	0.67	0.55	0.63	0.59	0.41

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. International Patent Classification: (1) *Human Necessities*, (2) *Performing, Operations and Transporting*, (3) *Chemistry and Metallurgy*, (4) *Textiles and Paper*, (5) *Fixed Construction*, (6) *Mechanical Engineering, Lighting*, (7) *Physics*, and (8) *Electricity*.

Table 2.34: By Technological Fields

This marked the beginning of a canal-building era, with a significant surge between 1815 and 1860, also called the "canal fever". These canals aimed to create an extensive network linking rivers and establishing efficient transportation routes across the country.

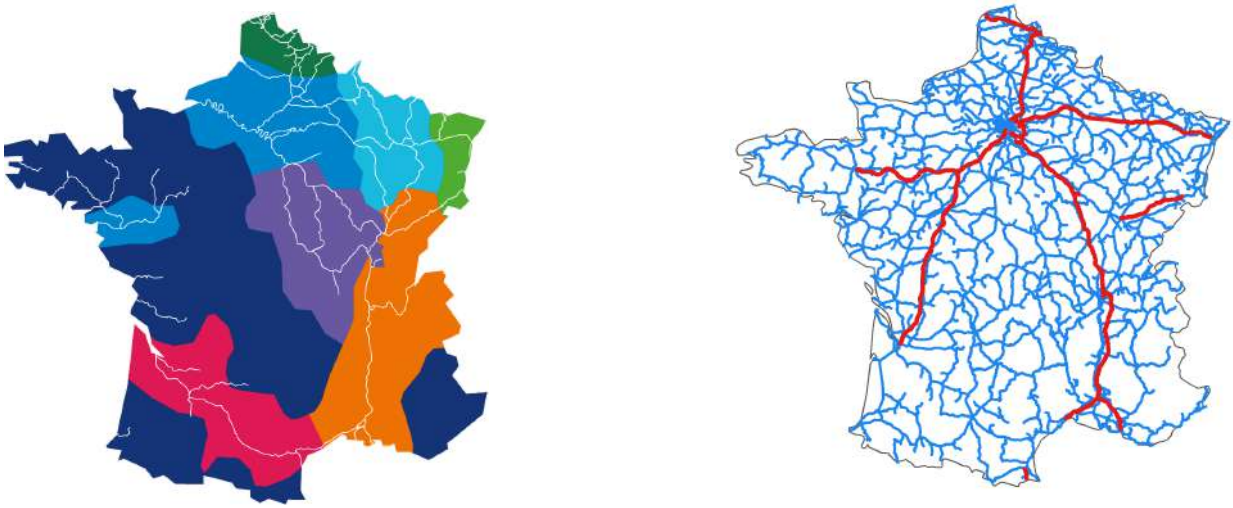
Canals played a crucial role in facilitating trade, providing cost-effective transportation for goods, and fueling economic development during the pre-industrial and industrial eras (19th century in France). Additionally, canals fostered regional integration by connecting different regions, promoting cultural interaction and economic cooperation. They also supported industrialization by facilitating the movement of raw materials and finished products. Furthermore, canals served agricultural purposes, providing irrigation for farmlands and improving agricultural productivity. Overall, the construction of canals in France had wide impacts on trade, regional integration, industrialization, and agricultural activities.

Since the canals helped establish and intensify trade routes, they impacted the interactions and economic relationships between the cities along their routes. These interactions have potentially continued over time and still exist today by historical anchoring. Hence, the

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
same$_{ij} \times t$	-0.01	-0.03^{***}	-0.03^{***}	-0.01	-0.03^{***}	-0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
contiguous$_{ij} \times t$	0.01^{**}	0.01^{**}	0.01^{*}	0.01^{**}	0.01^{**}	0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
log(distance$_{ij}$) $\times t$	0.01^{***}	0.00^{*}	0.00	0.01^{***}	0.00^{*}	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
log(travel time$_{ijt}$)	0.00	-0.12	-0.35			
	(0.12)	(0.13)	(0.22)			
log(travel time$_{ijt}$)$\times \mathbb{1}$(Core-Core)				-0.05	-0.18	-0.48^{*}
				(0.12)	(0.14)	(0.25)
log(travel time$_{ijt}$)$\times \mathbb{1}$(Core-Periphery)				0.51^{**}	0.25	-0.00
				(0.24)	(0.25)	(0.35)
log(travel time$_{ijt}$)$\times \mathbb{1}$(Periphery-Periphery)				0.24	0.36	0.71
				(1.00)	(1.03)	(1.62)
asinh(bridges)	0.15^{***}	0.17^{***}	0.18^{***}	0.15^{***}	0.17^{***}	0.18^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
asinh(technosim)	0.29^{*}	-0.10	-0.04	0.31^{*}	-0.08	-0.03
	(0.16)	(0.16)	(0.18)	(0.16)	(0.16)	(0.18)
Num. obs.	122714	115470	73945	122714	115470	73945
Num. groups: ij	4332	4148	2722	4332	4148	2722
Num. groups: it	2556	2535	2080	2556	2535	2080
Num. groups: jt	2556	2535	2080	2556	2535	2080
Deviance	72981.01	67222.61	41621.70	72970.36	67217.52	41619.01
Log Likelihood	-70685.04	-64418.89	-39097.72	-70679.72	-64416.35	-39096.37
Pseudo R^2	0.80	0.78	0.68	0.80	0.78	0.68

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are in parentheses. The dependant variable is the count of co-patents developed between region i and j, identified by the location of inventors residence. All columns include region-year fixed effects, γ_{it} and δ_{jt}, as well as pair fixed effects, ρ_{ij}. In order to control for the general increase in inter-regional co-patenting practices, we control for time trends in pairs where $i = j$, in contiguous pairs, and in distance. Results find that as time goes by, there are more co-patents developed for each distance value except within intra-regional borders. This effect erase the significance associated to travel time, except for *Core* regions not directly connected by HSR.

Table 2.35: Inter-regionalization time trends (2)



(a) Navigable waterways (VNF) (b) High-speed railways

Figure 2.12: Comparison of waterways with high-speed rail networks

presence of waterways could have influenced the location decisions for high-speed railways between cities. The presence of active water transportation routes can indicate areas of economic activity and trade hubs. High-speed railways may be established in close proximity to these waterways to facilitate the movement of goods and capitalize on the existing economic potential.

High-speed rail lines may be strategically located near waterways to take advantage of this existing transportation infrastructure during the construction phase of high-speed rail lines. Waterways could have been utilized to transport construction materials such as steel and machinery. Water transportation is often cost-effective for heavy and large materials, allowing for efficient supply to construction sites.

Alternatively, the presence of relatively flat terrain along waterways can facilitate the construction of high-speed railways. Moreover, the flat terrain along waterways provides favorable conditions for maintaining high-speed operations. The absence of significant inclines or curves allows trains to maintain their speed more efficiently, resulting in improved energy efficiency and reduced travel times.

Figure 2.12 motivates the argument developed above that the presence of waterways could have influenced the location decisions for high-speed railways. Figure 2.12a is a map showing the navigable waterways in France, diffused by *Voies navigables de France*, a public administrative institution responsible for the management of inland waterways network in France. Figure 4.1 shows the high-speed rail network today. Both images exhibit striking similarities in the way they connect various locations within the French territory. They share a dense coverage over the eastern side of France. On the western side, high-speed railways occupy areas where navigable waterways are absent, but where navigation through maritime routes is possible to connect the different locations.

2.C.2 Data

Inland navigable waterways : BD CARTO provides a vector description of various elements of the landscape with decametric (10 meters) precision, and in particular on the hydrographic

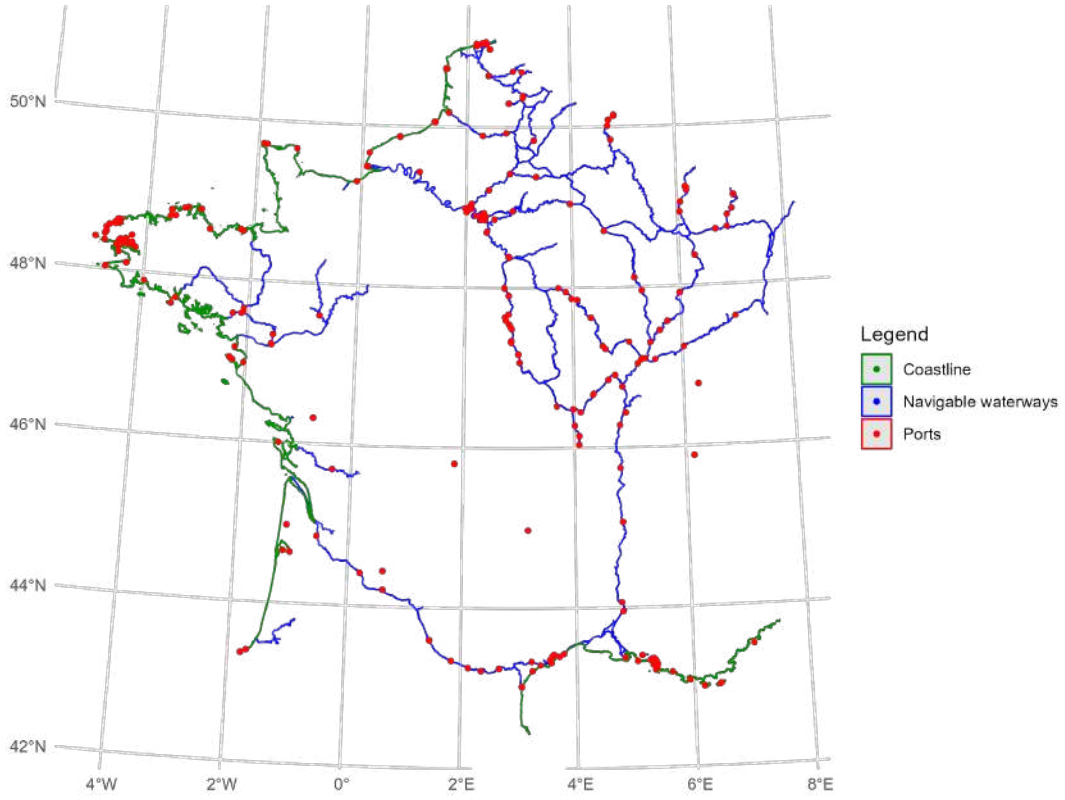


Figure 2.13: Navigable waterways

network. it provides a shapefile with the geographical location of waterways and an index which indicates wether the waterway is navigable or not.

Coastlines and ports : Open Street map, Overpass Turbo

2.C.3 Methodology

To build the instrument we follow the following steps:

1. Gather the shapefiles of inland navigable waterways and coastlines together.
2. Rasterize the new spatial object with a 50x50 cells grid (quite gross) - we obtain figure 2.14a.
3. Assign cells values to 1 for waterways and to 0 for land areas - we obtain figure 2.14b.
4. Same of ports - we obtain figure 2.14c.
5. Overlap the two rasters.
6. Create a network with cells being nodes and contiguity of cells being links - we obtain figure 2.14d (the inland and coastline waterways are shown in the background to show that the created network overlap the waterways network).
7. Run Djikstra algorithm to compute distance between every port.
8. Overlap cells location with sahpefile of NUTS3 regions in France.
9. For each pair of regions, take the average distance between their ports.
10. For pairs of regions that cannot be accessible via navigable waterways, the distance measure has NA values.

Then, the instrument takes 0 value when the distance by waterways is not computable or regards pairs involving same regions ($i = j$), and takes value 1 when the distance value is positive. In order to better fit the endogenous variable, we set value 0 for pairs with low distance.[30]

$$\text{waterways}_{ij} = \mathbb{1}(\text{waterways distance} \in \mathbb{R}) \times \mathbb{1}(i \neq j) \times \mathbb{1}(\text{contiguity} = 0) \tag{2.11}$$

2.C.4 Descriptive statistics

In our sample, 32% of all non-adjacent pairs of regions (with $i \neq j$) are connected by waterways. 77% of non-adjacent pairs connected by waterways experienced a decrease in travel time. On the other hand, 64% of the non-adjacent pairs of regions (with $i \neq j$) have experienced a decrease in travel time, of which 39% are connected by waterways.

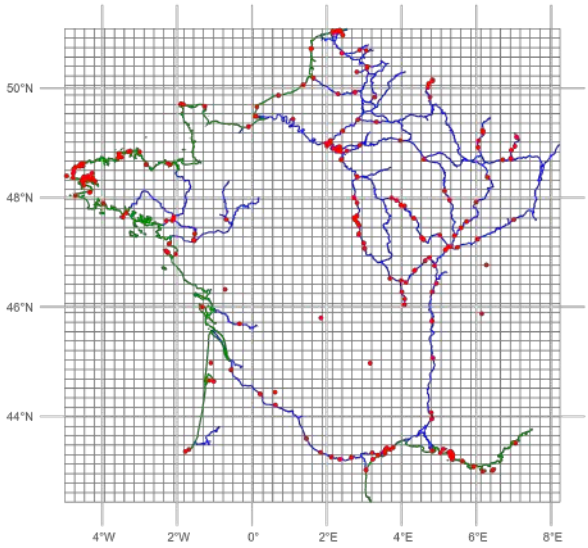
We estimate the following cross-section linear probability model:

$$\mathbb{1}(\Delta_{2010,1980} \log(\text{travel time}_{ijt}) < 0) = \alpha \text{ waterways}_{ij} + FE_i + FE_j + u_{ij} \tag{2.12}$$

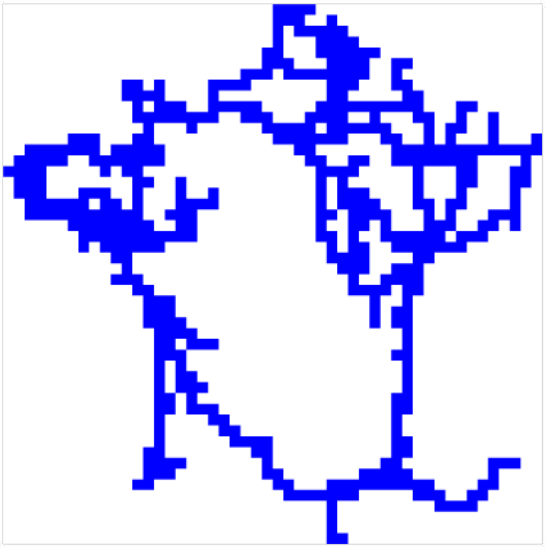
with $\Delta^{1980,2010} \log(\text{travel time}_{ijt})$ the change of travel time (in log) between 1980 and 2010.

Results displayed in figure 2.36 show that being connected by waterways (for non-adjacent regions) increases the probability to experience a decrease in travel time due to HSR roll-out by 18%.

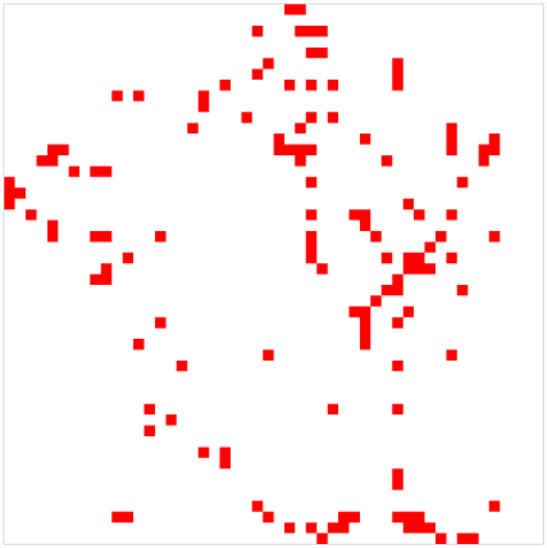
[30] Without this second restriction, there is even more endogeneity in the results.



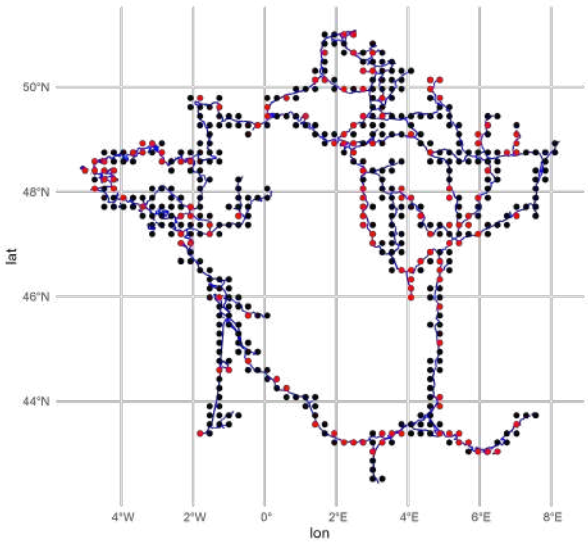
(a) Grid



(b) Waterways Raster



(c) Ports Raster



(d) Network

Figure 2.14: Illustration of the instrument computation

	$\mathbb{1}(\Delta_{2010,1980} \log(\text{travel time}_{ijt}) < 0)$
waterways_{ij}	0.18^{***}
	(0.02)
Num. obs.	8281
Num. groups: i FE	91
Num. groups: j FE	91
Adj. R^2 (full model)	0.33
Adj. R^2 (proj model)	0.01

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2.36: Linear Probability Model

Chapter 3

A Train Travel Time Dataset: Intercity and High-Speed Railways in France (1980-2020)

JOINT WITH FERNANDO STIPANICIC

Abstract

This paper presents a novel dataset of city-to-city travel time by train in France, covering the period 1980-2020, as well as the methodology we followed to construct it. We used an arrival-departure time schedule from Société Nationale des Chemins de Fer (SCNF), the French national state-owned railways company, as well as dates of high-speed railways (HSR) openings. From 1981 to 2017, high-speed lines have been built to connect Paris to major cities in France. Using Dijkstra algorithm (Dijkstra, 1959), we compute the contemporaneous travel time between every two cities in France. Then, to compute the past values of travel time within each pair, we rely on the assumption that prior to an HSR opening, trains were running at a normal speed. We are able to compare our estimations of travel time by train to observed values of travel time from a subsample of city-pairs (SNCF). Our dataset is found to replicate 95% of the overall variation in the observed travel time.

Keywords: High-Speed Railways, Travel Time, Transportation, Data
JEL Classification: C80, O18, R40

3.1 Introduction

In this paper, we present a novel dataset on travel time by train between every pair of cities in France from 1980 to 2020. France's rail transportation have evolved considerably since the introduction of the first high-speed railway (HSR) between Paris and Lyon in 1981. Eight major lines have opened since, until 2017 with the last two lines between Le Mans and Rennes, and between Tours and Bordeaux.

To do so, we rely on trains schedule datasets from December 2021 made available by the Société Nationale des Chemins de Fer (SNCF), France's national state-owned railways company. From this data, we are able to estimate the train travel time between every two cities in France using the Dijsktra algorithm, which estimate the shortest path in terms of travel time between every pair of stations.

To estimate past values of travel time, we identify each station-pair connected by an HSR, as well as the year of the roll-out, and rely on the assumption that prior to an HSR roll-out, trains were running at a normal speed between two stations. The normal speed of trains is estimated according to the 2021's train speed on normal (non-high-speed) lines. Since the information we use is limited to the contemporaneous rail network and the history of high-speed railways, the travel time variation that we estimate is only due to the opening of high-speed lines.

Despite the fact that our estimated travel time between cities rely on simplified assumptions, it is found to be very close to what can be observed for a small sample of city pairs (SNCF). The observed travel time variation is accounted for about 95% by our estimated travel time. Our dataset is the first available dataset on the evolution of travel time between every pair of cities in France.

The paper is organized as follows. Section 3.2 presents the raw data of SNCF that we use for our estimations of travel time. Section 3.3 presents the history of high-speed railways in France. Section 3.4 presents our methodology we follow to compute our dataset. Section 3.5 evaluates the validity of our estimations. Section 3.6 presents our final dataset. Finally, section 3.7 concludes.

3.2 Arrival and departure time schedule

SNCF made available schedule datasets reporting the departure and arrival time for all non-stop train services currently realized. Datasets are split by train type: TGV, Intercités, TER and Transilien. TGV (Train à Grande Vitesse) corresponds to high-speed trains, Intercités are trains operating between cities on the main train lines at a normal speed, while TER (Transport Express Régional) and Transilien are trains operating on regional lines, which have a more local use. TGVs operate on high-speed railways as well as on normal lines, while the other trains only run on normal lines. We download the datasets available in December 2021, which include all trips undertaken by trains between the 8th and 16th December 2021.

The raw schedules includes four datasets for each of the four train types, resulting in 16 different datasets. A first dataset **`routes.txt`** gives an identifier for each unique route undertaken by trains from the start to the end main station, as well as their name. For example, we find 321 different routes in the TER dataset,[1] such as *Rennes-Brest* or *Paris-Dijon-Mulhouse*.

[1]The dataset also includes routes undertaken by bus. We erase this information from our dataset since we

Train type	Route ID	Trip ID	Stop ID	Stop name
TGV	31	4,786	523	197
Intercités	11	1,059	282	135
Transilien	53	22,874	2,070	962
TER	321	18,579	2,484	2,472

Notes: Route ID corresponds to the route identifier, associated to a starting and final station. Trip ID is the trip identifier, which is unique depending on the time and direction of the trip. Stop name is the name of the station, and the stop ID is the station platform group identifier. For large stations, we can count several stop ID since they count several groups of platforms. TGV are high-speed trains, running on both high-speed and normal railways. Intercités, Transilien and TER run on normal lines only.

Table 3.1: Train time schedule datasets (SNCF) - Amount of observations

A second dataset **`trips.txt`** gives information on the route and the trip identifier. While the former identifies the route whatever the time of day, the latter is unique depending on the time of the day the service operates from the first to the last station of the route. For example, the TER route *Rennes-Brest* counts 47 different trips. One trip among them is the trip departing from Rennes at 6.02am and arriving in Brest at 8.20am. Another trip runs in the opposite direction and departes from Brest at 1.08pm to arrive in Rennes at 3.25pm. This time information is available in **`time.txt`**, which contains the trip identifier, the station identifier, as well as the arrival and departure time of the train at each station. Finally, **`stops.txt`** gives information on the station identifier, the station name and its latitude and longitude coordinates. For the *Rennes-Brest* trip departing from Brest at 1.08pm, we are able to observe that the train passes through 9 stations in the corresponding order: *Brest, Landerneau, Landivisiau, Morlaix, Plouaret Trégor, Guincamp, Saint-Brieux, Lamballe, Rennes*, and at what time the train arrives and leaves each station.

Table 3.1 summarizes the amount of observations by the main identifiers for each train type. The train type that operates on the highest amount of routes is the TER, since it operates intra-regionally in the whole France except Île-de-France, Paris' region, where operate Transiliens. TER pass through almost 2,500 different stations in France. Intercités and TGV count a lower amount of different routes since they operate at large distances within the country.

The goal of the present work is to estimate the travel time between every two cities in France for each year since the opening of the first high-speed railways. The datasets presented in this section enable us to compute a contemporaneous travel time between stations.[2] To be able to estimate the past values of travel time, we need to identify the pairs that are treated by an HSR, as well as the year of their opening. Next section presents the history of the French high-speed railways.

3.3 History of the French high-speed railways

France inaugurated its first high-speed railways in 1981 between Lyon and Saint Florentin,[3] which intends to connect Paris, the capital city, and Lyon, the second main french city. Since

want to estimate the train travel time.

[2]The possibility to change stations within a city is not integrated in the dataset since they are usually not located on the same railway. Thus, we compute a fictive travel time between stations within a same city. Section 3.4 explains the computation.

[3]Saint Florentin is located at 300km to the North-East of Lyon and at 150km to the South-West of Paris.

then, the network has expanded considerably, connecting Paris to the biggest cities in France. Figure 3.1 illustrates the high-speed railways expansion over time by showing the rail network by decade from 1990 to 2020. The high-speed network is shown to display a star shape, with the capital city Paris at its center. High-speed trains can reach a maximal speed of 320km/h, compared to the less than 160km/h for Intercités. By the end of 2017, we count more than 1,500km of high-speed railways. We present below a list of the dates in which the different high-speed railways (Ligne à Grande Vitesse - LGV) opened:

- 1981: First part of the line LGV Sud-Est aiming at connecting Lyon to Paris. The line opened on May 22nd between Saint Florentin and Lyon. Saint Florentin is around halfway between Paris and Lyon.
- 1983: Second part of the line LGV Sud-Est reaching Paris opens on April 25th.
- 1989: First branch of the line LGV Atlantique, Paris - Courtalain - Connerré - Le Mans.
- 1990: Second branch of the line LGV Atlantique, Courtalain - Saint-Pierre-des-Corps - Monts.
- 1991: Opening of Massy station.
- 1992: East bypass of Lyon from Montanay to Saint-Quentin-Fallavier (in order to create the line LGV Rhône Alpes, which would connect LGV Sud-Est to the extreme south of France) in December 1992.
- 1993: LGV Nord from Paris Gare du Nord to Lilles-Flandres, opened on May 18th.
- 1993: LGV Nord from Lille-Europe to Calais-Fréthun (towards the Channel Tunnel which connects France to England), opened on September 26th.
- 1994: LGV Rhône-Alpes, from Lyon-Saint-Exupéry to Valence.
- 1994: East bypass of Paris Interconnexion Est at Vémars-Couvert-Crisenoy (connects Aéroport Charles de Gaulle and Marne-la-Vallée Chessy to the LGV).
- 1996: East bypass of Paris Interconnexion Est at Valenton-Coubert.
- 1997: Lille-Flandres to Belgium border towards Bruxelles, opened on December 10th.
- 2001: LGV Méditerranée connecting Marseille to Paris via LGV Sud-Est (Paris-Lyon), LGV Rhône-Alpes (Lyon-Valence) and LGV Méditerranée (Valence-Marseille). Fork towards Avignon and Nîmes.
- 2007: First part of the line LGV Est from Paris-Est to Baudrecourt, opened on June 10th.
- 2010: Line towards Spain, Perpignan-Figuières, opened in December 19th.
- 2011: First part of LGV Rhin-Rhône from Villiers-les-Pots to Petit-Croix, opened on December 11th.
- 2016: Second part of the line LGV Est between Baudrecourt and Vendenheim, opened on July 3rd.

- 2017: LGV Sud Atlantique between Tours and Bordeaux, opened on July 2nd (extension of the second branch of the line LGV Atlantique which connects Paris to Tours).
- 2017: LGV Bretagne-Pays de la Loire, which connects Le Mans to Rennes, opened on July 2nd (extension of the first branch of the line LGV Atlantique which connects Paris to Le Mans).
- 2017: Bypass of Nîmes and Montpellier, LGV Méditerranée, opened on December 10th.

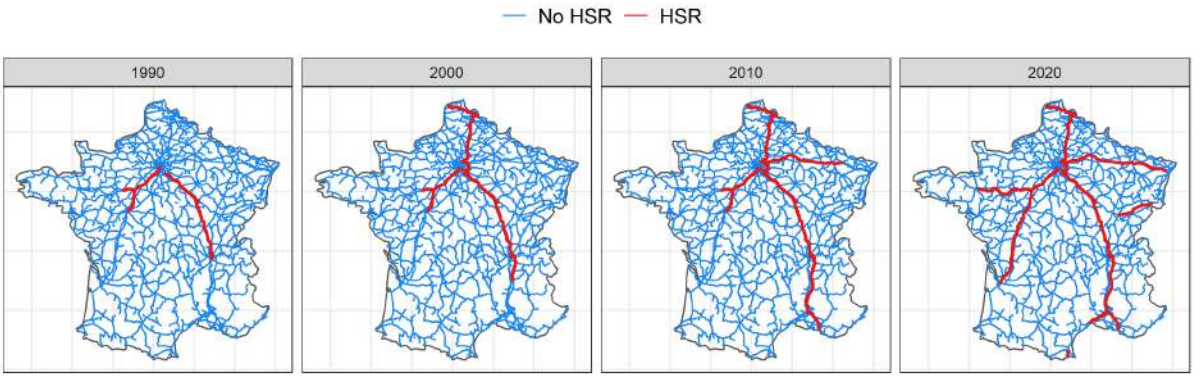


Figure 3.1: High-speed railways expansion by decade

From this historical information, we manually identify the non-stop stations pairs (from the raw SNCF datasets) connected by an HSR and the year of opening. Some pairs are connected by both HSR and normal lines (see the example of *Paris - Chambéry* pair in section 3.4). Hence, we also differentiate pairs that are fully treated by an HSR and those that are not fully treated. All information needed in order to compute the travel time evolution between every pair of cities is gathered. Next section presents our methodology.

3.4 Travel time computation

We gather all the 16 datasets presented in section 3.2 in order to have a complete schedule of trains operating in metropolitan France, with information on the route, the trip, the stations, as well as the train arrival and departure time in each row of the dataset. Then, we delete trips which end before 6.00am and start after 9.00pm in order to consistently compute travel time between non-stop stations pairs.[4] We merge this schedule with itself by route and trip ID, in order to get pairs of stations for each single trip. Then, we keep the pairs in which a train runs without any stop in the between, i.e. the direct connections only, as well as the departure time from the origin station and the arrival time to the destination station. This station-to-station dataset allows us to compute the observed travel time between every two stations with a non-stop connection for each train service by computing the difference between the departure time in the origin station and the arrival time in the destination station. We keep

[4]Most of train trips occur between 6.00am and 9.00pm.

the minimum value of travel time between each non-stop stations pair to keep pairs' unique information.

At this stage, the output table is still looking raw, but will allow us to compute an estimated contemporaneous travel time between every two stations in France. Figure 3.2 presents the observed travel time for each non-stop stations pair by train type relative to the geodesic distance between the stations. TGV trains are separated into three groups depending on the type of railways they run on within each pair of stations: high-speed railways, normal railways or a mix of both. The identification of high-speed railways within pairs is described in section 3.3. Out of the 34,955 municipalities in France, we count 2,498 municipalities served by trains in the dataset. Among them, 164 are crossed by high-speed trains - since they run on both high-speed and normal lines - and 40 are directly connected to high-speed rail network - where trains run at high-speed, i.e. about 200km/h on average.

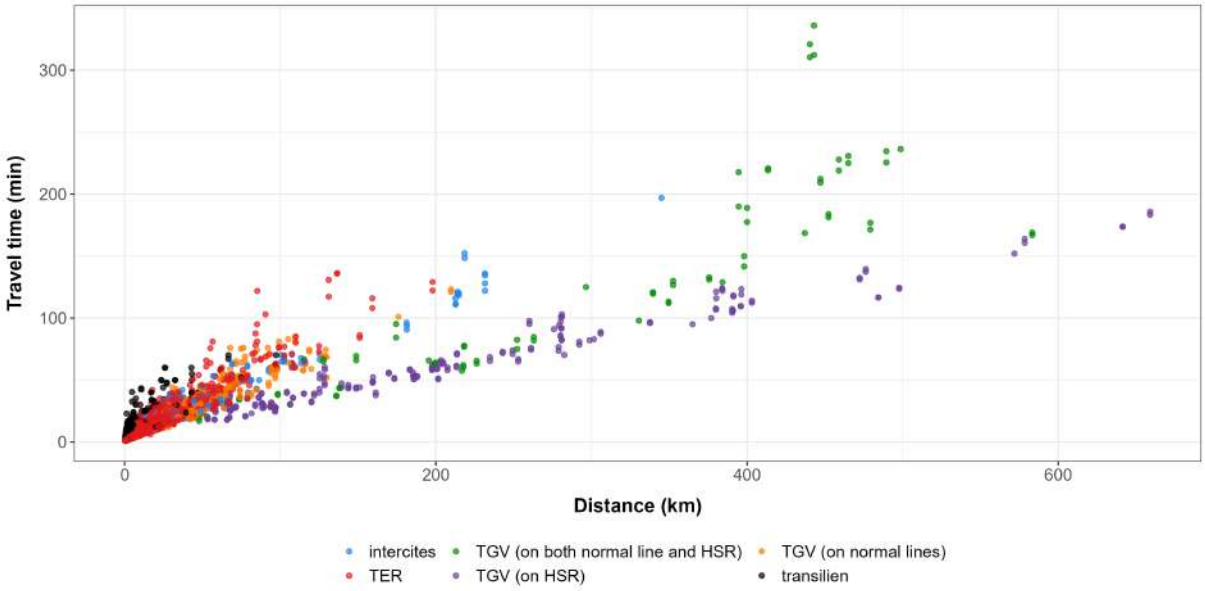


Figure 3.2: Observed travel time and distance within non-stop station pairs by train type

TERs and Transiliens are found to operate on much lower distances between stations than the other trains - see Table 3.2 for a description of the distance within stations pairs. Stations pairs crossed by TERs and Transiliens are found to be around a fictive 45 degree line on the plot - which translates 1 minute for 1 kilometer. Stations pairs crossed by Intercités and TGV are found to be below the 45 degree line. These trains run through longer distances at a lower travel time by kilometer, i.e. at a higher speed. The relationship between travel time and distance looks particularly linear for each transport mode. However, it is not the case for non-stop pairs in which TGV operates on both HSR and normal lines. In the schedule dataset, some routes show a station pair to be direct for some trips but not direct for others. For example, the pair *Paris - Chambéry*[5], which represents 450km of distance, is found to be direct for the trip leaving Paris at 7.35am and arriving in Chambéry at 10.58am, but counts a stop in Lyon in the between for the trip leaving Paris at 6.39am and arriving in Chambéry at 9.39am. The pair *Paris - Lyon* is connected by an HSR all the way long, but the pair *Lyon - Chambéry* is connected by a normal line. Hence, while the pair *Paris - Chambéry* is direct, it is not fully

[5]Chambéry is located at a hundred of kilometers to the west of Lyon, in the French Alps.

connected by an HSR. For those pairs, travel time is increasing more and more as distance increases.

Table 3.2 presents the descriptive statistics of distance between stations. Among the TER travels, trains run from one station to the other between 0.3km to 14km without stop in the between for 75% of the stations pairs. It is even lower for Transilien, which count 75% of the stations pairs with a distance between 0.4km and 5km between stations. This result is consistent to the fact that Paris' region, in which operate Transilien, displays a strong population concentration, much higher than the other regions in France. As a consequence, trains stop more frequently and therefore at shorter distances. The average distance within pairs crossed by Intercités and TGV on normal lines is about 58km and 41km respectively. It is much higher than for Transilien and TER since the two former trains are aimed for inter-regional transportation, while the two latter are aimed for a more local use.

	min	1st qu.	median	mean	3rd qu.	max	obs
TGV (on HSR)	21.48	94.27	183.23	209.18	280.65	659.08	197
TGV (on normal lines)	2.35	20.13	33.38	41.62	55.34	209.77	378
TGV (on HSR and normal lines)	17.73	137.97	296.50	286.22	437.04	583.21	77
Intercités	2.94	22.51	39.29	58.29	68.82	344.94	209
Transilien	0.37	1.64	2.68	5.16	4.93	97.36	1,637
TER	0.30	4.22	7.52	11.65	14.00	197.94	6,793

Table 3.2: Descriptive statistics - distance (km) between non-stop stations pairs

	min	1st qu.	median	mean	3rd qu.	max	obs
TGV (on HSR)	76.98	173.67	197.34	189.26	211.82	255.90	197
TGV (on normal lines)	18.73	69.19	89.84	87.93	105.77	149.74	378
TGV (on HSR and normal lines)	79.11	124.61	155.41	153.54	185.24	226.09	77
Intercités	22.04	66.75	86.09	85.11	105.98	135.09	209
Transilien	3.24	31.97	47.47	46.40	60.63	120.53	1,637
TER	9.40	55.45	70.61	70.91	85.27	196.35	6,793

Table 3.3: Descriptive statistics - running speed (km/h) between non-stop stations pairs

The travel time - distance relationship can be observed through the speed at which trains travel. Figure 3.3 presents descriptive statistics on running speed by train type between non-stop stations pairs. For each pair, the speed expressed in kilometers per hour is computed as follows: speed $= 60 \times$ distance/travel time, with the distance expressed in kilometers, the travel time expressed in minutes. We find that the average speed of TGV is much higher than the other trains. On high-speed railways, it runs at 189km/h in average, up to 256km/h for a non-stop pair, i.e. *Lorraine TGV - Champagne-Ardennes TGV*. High-speed trains running on normal lines and Intercités are comparable. They run at a speed around 85km/h in average. TER and Transilien run at 71 and 47km/h respectively, consistent with the fact that they operate more locally than the other trains. Regarding maximum values of speed, we find that the statistic for TER is unexpectedly high. Indeed, the difference between the 3rd quartile and the maximum value is substantial. Investigating the data, we find that some pairs of TER travels correspond to TGV travels, and most importantly, to TGV travels on high-speed lines.

Nowadays, TER may be able to run faster on appropriate railways. However, to be consistent with the history of HSR openings, we erase these observations from the dataset for the next computations.

In order to estimate the travel time past values, we rely on three information: (1) the opening of HSR, (2) the contemporaneous average speed of inter-regional trains, i.e. Intercités, since it is comparable to TGV running on normal lines, and (3) the geodesic distance between each non-stop stations pair, which is used as an approximation for the rail distance. For stations pairs with a high-speed railway in the between in December 2021, we uses the contemporaneous observed travel time after an high-speed railways (HSR) has been introduced and we estimate the travel time before the HSR roll-out using contemporaneous Intercités average speed and station pairs' distance. For the years before the introduction of an HSR, we erase from the datasets the pairs that are connected by an HSR but not on the whole route between them, i.e. the partially treated pairs - see the example of *Paris - Chambéry* pair. Hence, the direct connections remains after the HSR roll-out even if not fully treated, and prior to the HSR connection, segments will be split between full-HSR and full-normal lines. For all other stations pairs, we use the contemporaneous observed travel time.

To compute the estimations of past values of travel time, we estimate the average effect of distance on travel time between stations. We estimate the following regression:

$$\text{travel time}_{od\tau} = \alpha_{0\tau} + \alpha_{1\tau}\text{distance}_{od\tau} + u_{od\tau} \tag{3.1}$$

with *od* the non-stop pair between the origin and the destination stations and τ the train type. Travel time is expressed in minutes and distance is expressed in kilometers. Table 3.4 shows the results. We find that TGV operating on HSR display a lower coefficient than the other trains. For 10 additional kilometers between two stations, the TGV travel time is expected to increase by 3 minutes, while the Intercités travel time is expected to increase by 5 minutes in average. The same result applies for TGV running on normal lines. From these coefficients, we compute the implied average cruise speed as follows: $\widehat{\text{speed}}_{\tau} = 60/\hat{\alpha}_{1\tau}$. Intercités are found to run at an average speed about 111km/h, while we find an average speed of 229km/h for the TGV operating on HSR.

Following these results, the non-stop stations pair dataset is augmented with the estimated travel time between non-stop stations pairs prior to introducing an HSR between them. The estimated travel time for those fully-treated pairs is computed as follows: $\widehat{\text{travel time}}_{od,\text{ before HSR}} = 6.54 + 0.54 \times \text{distance}_{od}$, with the coefficients being those found for Intercités.

In order to have a complete train network and include the possibility to change stations within a city, we also estimate the intra-city travel time between stations. 11.85% of the cities in the dataset are found to have more than one station. We pair each station within cities and compute the geodesic distance between them. We find that Marseille displays the maximum distance between its stations, with a distance about 16km between the stations of L'Estaque and La Barasse. The average distance within cities is found to be about 4km and the third quartile is about 6km. Then, we compute the average speed of trains for non-stop stations within a distance of maximum 16km. Among the trains operating more locally, we find that transilien runs at a speed about 45km/h, while TER run about 67km/h. Since people are likely to take a bus, taxi, metro or tramway in order to change stations if they need to, they are likely to go at the same speed than transilien. We compute the travel time between stations within cities as follows: $\widehat{\text{travel time}}_{od,\text{ intra-city}} = \overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}} \times \text{distance}_{od}$, with $\overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}}$

Dependent Variable:	observed travel time				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
constant	6.54^{***}	8.80^{***}	3.91^{***}	2.60^{***}	1.48^{***}
	(0.67)	(0.93)	(0.60)	(0.13)	(0.05)
distance	0.54^{***}	0.26^{***}	0.57^{***}	0.80^{***}	0.67^{***}
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)
Train type	Intercités	TGV_{HSR}	$TGV_{non\text{-}HSR}$	Transilien	TER
Implied average speed (km/h)	112	229	105	75	90
Fit statistics					
R^2	0.95	0.96	0.87	0.69	0.88
Num. obs.	209	197	378	1,637	6,793

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. The implied average speed is computed as follows: speed $= 60/\hat{\alpha}_1$ since the travel time is expressed in minutes and the distance in kilometers.

Table 3.4: Station-pair regressions of travel time on distance by train type

the average speed found for Transilien non-stop stations pairs separated by less than 16 kilometers[6]. We add those new intra-city pairs and estimated travel time in our schedule dataset.

Finally, relying on the travel time information for each non-stop station-pair, we run the Dijkstra algorithm (Dijkstra, 1959) to find the shortest path in terms of travel time between each pair of stations in France for the year 2020. Then, going backward in time, before the opening of an high-speed line between two stations, we replace the high-speed travel time value by the estimated travel time computed using coefficients from regression 3.1 and the distance between those stations, i.e. we compute $\widehat{\text{travel time}}_{od} = \widehat{\alpha}_{0,\text{intercités}} + \widehat{\alpha}_{1,\text{intercités}} \times \text{distance}_{od} = 6.54 + 0.54 \times \text{distance}_{od}$. In other words, in the years prior to introducing HSR, high speed is replaced by normal speed for each station-pair concerned.. Djikstra algorithm is executed 41 times for each year from 2020 to 1980, considering changes in speed on railways. Therefore, the output data set reports travel time variations which come from the roll-out of high-speed railways only. Since a city can have more than one station, some city-pairs display more than one value of travel time per year. In order to keep a unique information for a city-pair and year, we keep the minimum travel time for each. Next section presents the output dataset.

3.5 Validation exercise

SNCF made available a dataset of train travel time between a subsample of cities in France from 1920 to 2020 (see figure 3.3). Therefore, we are able to compare our estimated travel time with observed values over the years, in order to evaluate the relevance of our dataset. The SNCF subsample counts 43 different city-pairs and 37 different cities, with most of the pairs involving Paris. We merge this dataset with ours and estimate the following regression:

$$\text{observed travel time}_{odt} = \beta_1 \text{ estimated travel time}_{odt} + \delta_{od} + \delta_{ot} + \delta_{dt} + \varepsilon_{odt} \tag{3.2}$$

[6] $\overline{\text{speed}}_{\text{Transilien, dist}\leq 16\text{km}} = 0.75$, which equals 45km/h divided by 60 since travel time is expressed in minutes.

with *observed travel time*$_{odt}$ the travel time between origin-city o and destination-city d at year t from the SNCF subsample, *estimated travel time*$_{odt}$ our estimated travel time, δ_{od} an origin-destination city-pair fixed effect, δ_{ot} a origin-city-year fixed effect, and δ_{dt} a destination-city-year fixed effect. Table 3.5 displays the results.

Figure 3.3: SNCF subsample

Results show that our estimated travel time accounts for about 95% of the variation in the observed travel time from the SNCF subsample. Considering travel time change within pairs, the R^2 is still substantially high, i.e. about 80% (see column 2). The R^2 statistic is lower since the change of our estimated travel time over the years only comes from the high-speed railways openings. However, in reality, travel time have slightly changed throughout the years even without any HSR opening because of possible replacement of railways or trains with more efficient ones. The model without any fixed effects (column 1) indicates that we overestimate the travel time between cities by about 10 minutes in average, i.e. the constant is equal to -10.36. The estimated effect of the estimated travel time on the observed travel time $\hat{\beta}_1$ is around 1, meaning that a one-minute change in the former variable corresponds approximately to a one-minute change in the latter. This result holds across and within pairs, controlling or not for each city-year.[7]

As a complementary validation exercise, we present different plots showing the observed and estimated travel times for some pairs of cities, which belongs to the SNCF subsample of the evolution of train travel time from the beginning of the 20^{th} century. Figures 3.4 to 3.7 show the results. Figure 3.4 shows that the estimation of travel time closely follows observed travel time within the pairs between Paris, Lyon and Lille in variation as well as in levels. The same applies for the pairs involving Paris, Strasbourg, Le Mans and Nantes as showed in Figure 3.7. Note that each drop in travel time in the estimated travel time is due to an HSR opening. We observe that for these pairs, the drop in observed travel time also mainly come from the HSR opening.

[7] A t-test is performed to evaluate whether if the estimated coefficients are significantly different from one. Results show that they are. Still, we consider that the difference is low.

Dependent Variable:	observed travel time			
Model:	(1)	(2)	(3)	(4)
Variables				
constant	-10.36**			
	(5.015)			
estimated travel time	1.084***	1.126***	1.345***	1.152***
	(0.0363)	(0.0370)	(0.4641)	(0.3041)
Fixed-effects				
origin-destination		Yes		Yes
origin-year			Yes	Yes
destination-year			Yes	Yes
Fit statistics				
Observations	1,584	1,584	1,584	1,584
R^2	0.951	0.980	0.992	0.997
Within R^2		0.796	0.657	0.346
T-test $\{H_0 : \beta_1 = 1, H_1 : \beta_1 \neq 1\}$				
t-statistic	2.314	3.405	0.743	0.500
reject H_0	Yes	Yes	No	No
significance level	5%	1%		

Clustered (origin-destination) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3.5: Validation exercise - observed and estimated train travel time between major cities

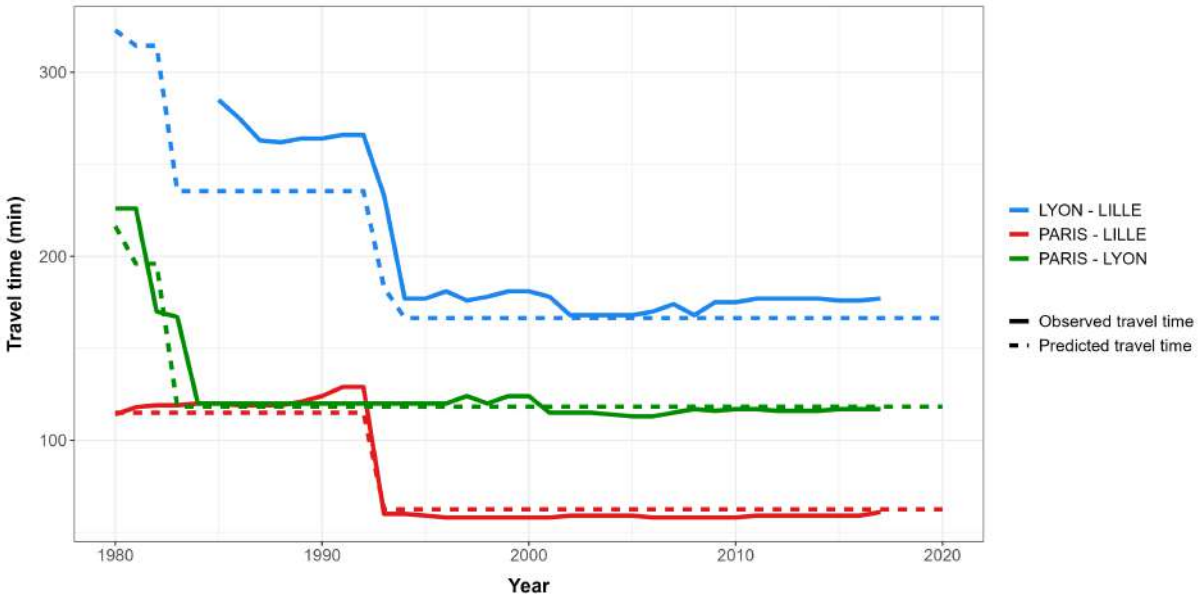


Figure 3.4: Observed and estimated travel time - Paris, Lyon, Lille

Figure 3.5 shows the evolution of travel time between Paris, Lyon and Bordeaux. In this case, we see that we overestimate the travel time between Paris and Bordeaux, probably be-

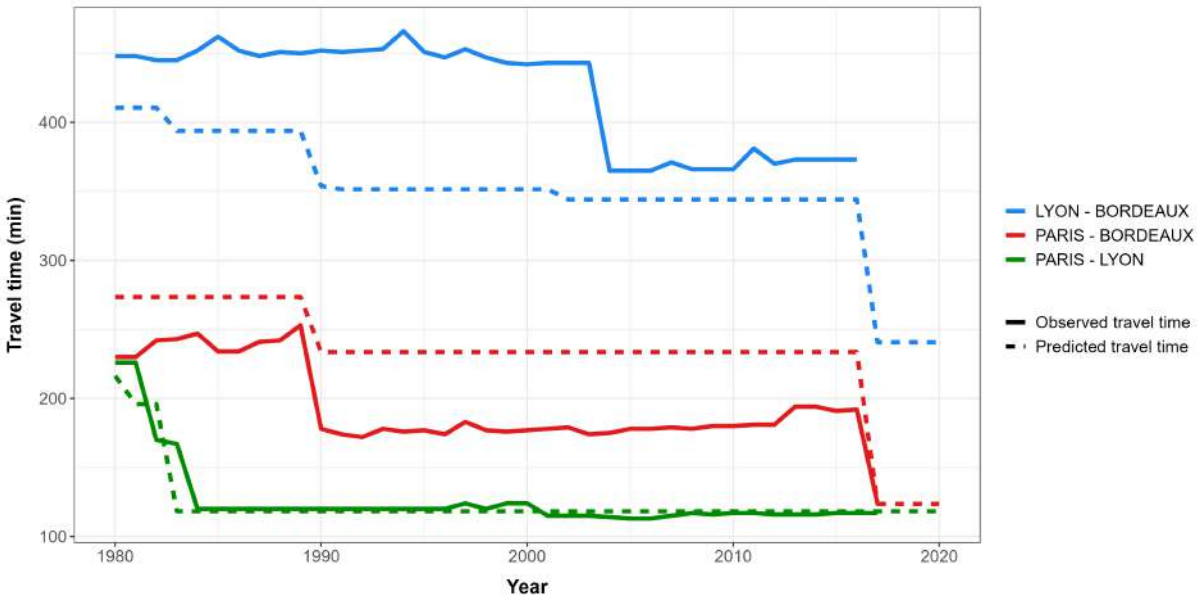


Figure 3.5: Observed and estimated travel time - Paris, Lyon, Bordeaux

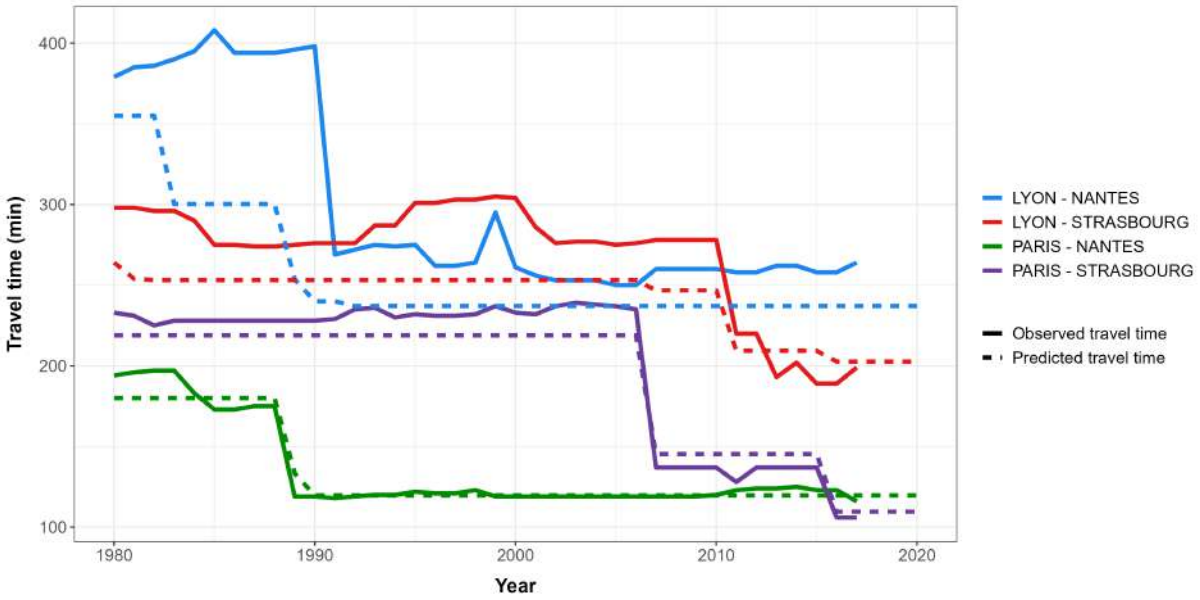


Figure 3.6: Observed and estimated travel time - Paris, Lyon, Nantes, Strasbourg

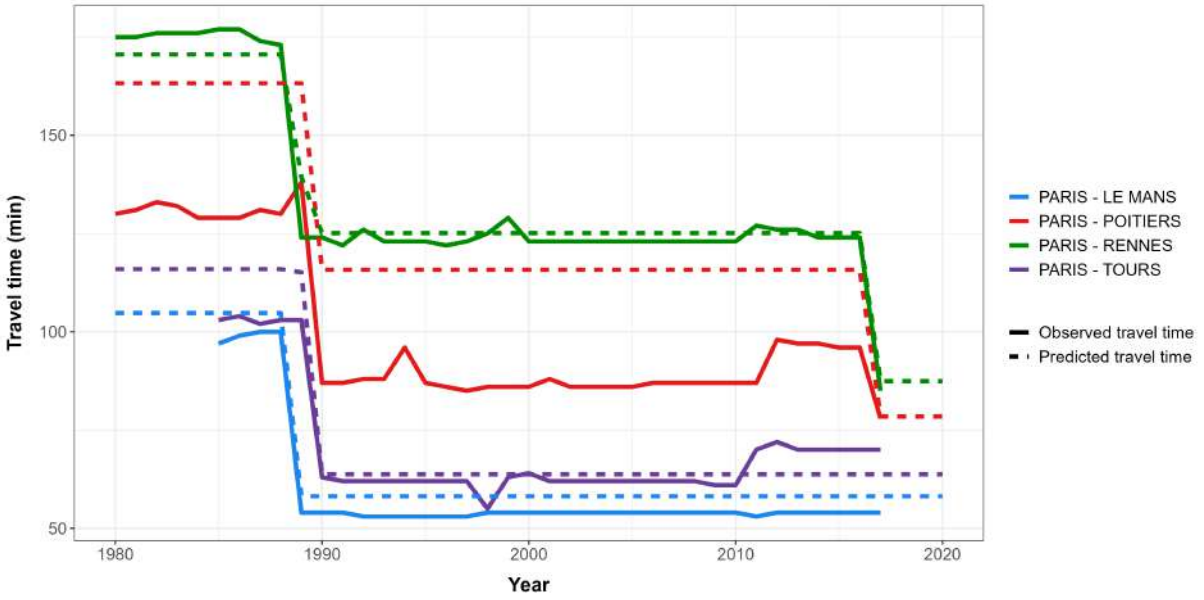
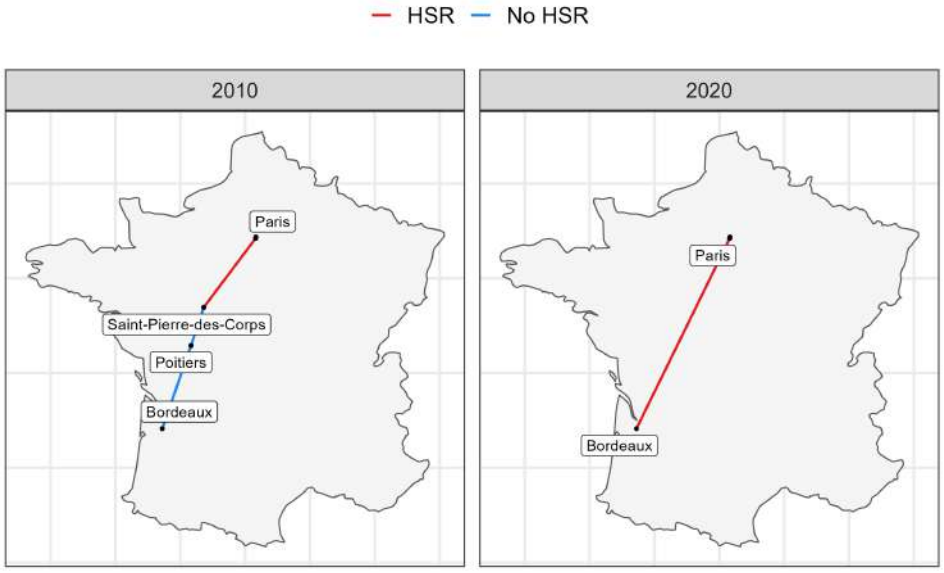


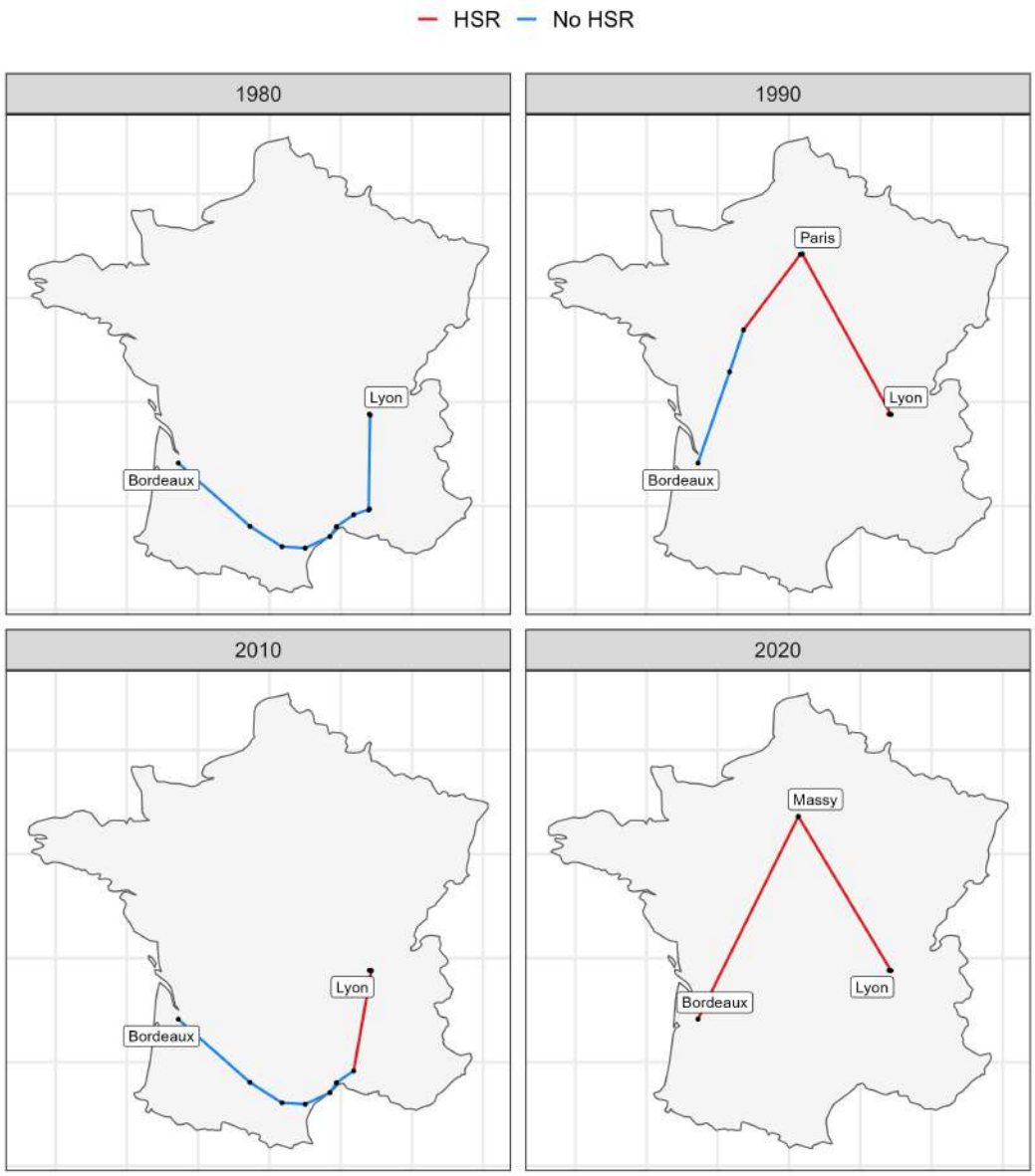
Figure 3.7: Observed and estimated travel time - Paris, Le Mans, Rennes, Tours, Poitiers



The figure shows the raw geodesic path undertaken by trains between Paris and Bordeaux in 2010 and 2020. The high-speed line between Paris and Tours (Saint-Pierre-des-Corps where is the station) has been opened in 1990. The second segment from Tours to Bordeaux has been opened in 2017. We recognize a direct trip from Paris and Bordeaux after 2017 since it is included in the original schedule dataset (SNCF). However, before the second HSR segment roll-out, the trip is non-direct since the whole way is split between high-speed and normal lines. We identify two stops in the between from the TER/TGV original schedule dataset (SNCF).

Figure 3.8: Shortest path Paris-Bordeaux

cause of long haul economies. Paris-Bordeaux pair has been treated by an HSR twice. The first HSR connection opened in 1990 between Paris and Tours and the second opened in 2017 between Tours and Bordeaux. Our computation assumes that prior to the HSR openings, non-stop connection between Paris and Bordeaux does not exist, but trains from Paris first stop

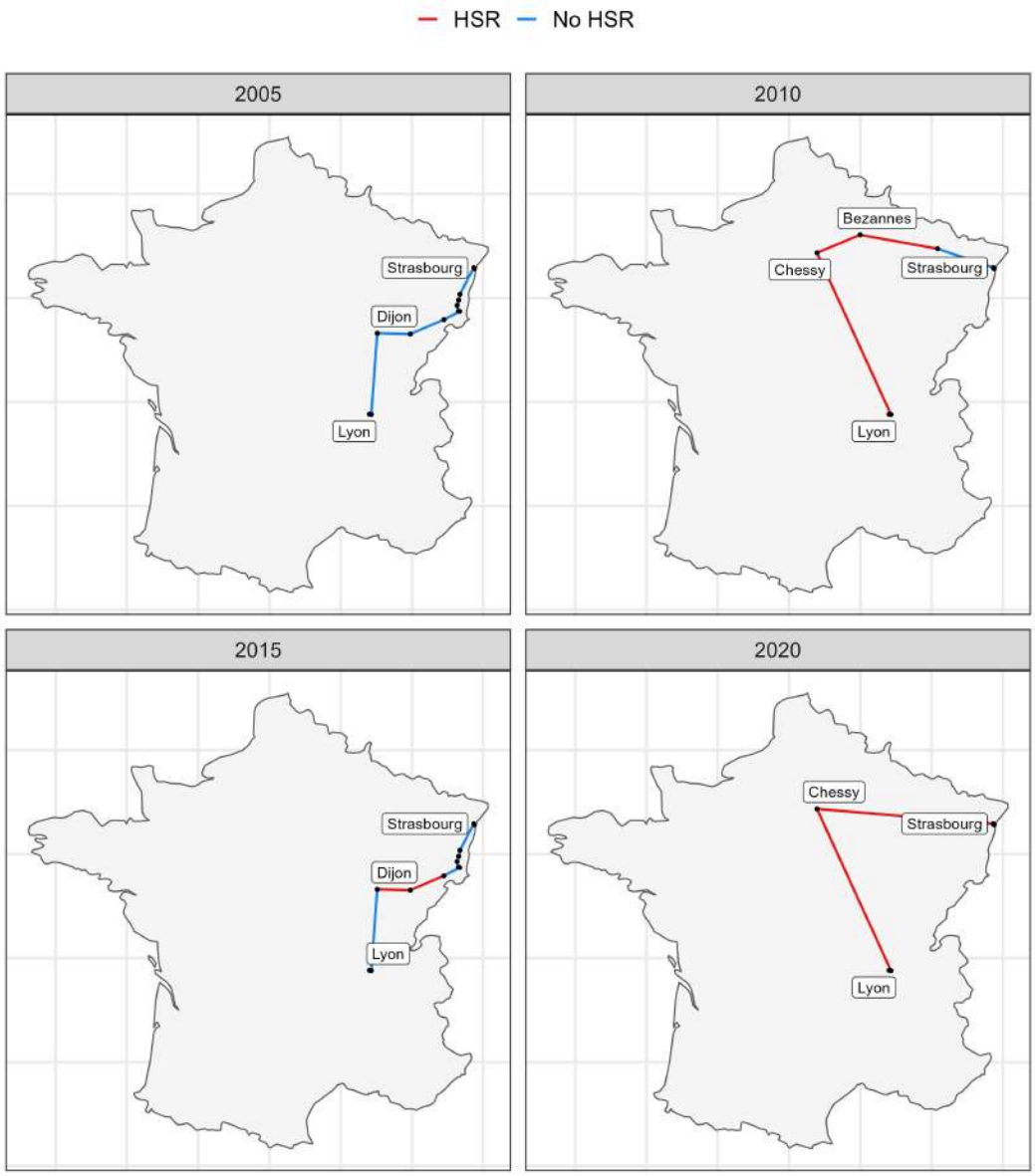


The figure shows the raw geodesic path undertaken by trains between Lyon and Bordeaux in 1980, 1990, 2010 and 2020. The high-speed line LGV Sud-Est between Paris and Lyon has been opened in 1983. In 1990, Paris and Tours are connected by the LGV Atlantique. In 1991, the high-speed train Massy station has opened, located about 20 kilometers from Paris. In 2001, LGV Méditerranée has connected Lyon to Marseille, with a fork towards Avignon and Nîmes on the line. Finally, in 2017, the second segment of LGV Atlantique from Tours to Bordeaux has opened in 2017. According to Dijkstra algorithm, the path undertaken by trains should have changed following the high-speed rails roll-out considering the minimum travel time to go from Lyon to Bordeaux each year.

Figure 3.9: Shortest path Lyon-Bordeaux

in Tours and Poitiers, then go to Bordeaux - and vice-versa (see figure 3.8). Doing so decreases the average speed of the train since it has to stop and start again - it is captured by estimated intercepts from regression 3.1. That could partially explain the average difference of 50 minutes between the estimated and the observed values of travel time. The same applies for Paris-Poitiers pair in Figure 3.7.

On the contrary, we see that we underestimate the travel time between Lyon and Stras-



The figure shows the raw geodesic path undertaken by trains between Lyon and Strasbourg in 2005, 2010, 2015 and 2020. The high-speed line LGV Est between Paris and Baudrecourt has been opened in 2007. The path is showed to go through Chessy station, about 30 kilometers from Paris, since it is located on the LGV Interconnexion Est opened in 1994. The second segment between Baudrecourt and Strasbourg has opened in 2016. In the meanwhile, LGV Rhin-Rhône has opened in 2011 between Dijon and Belfort-Monbéliar. According to the Dijkstra algortihm, the path undertaken by trains should have changed following the high-speed rails roll-out considering the minimum travel time to go from Lyon to Strasbourg each year.

Figure 3.10: Shortest path Lyon-Strasbourg

bourg (figure 3.6) or between Bordeaux and Lyon (figure 3.5) for example. In this case, we face another problem which may come from the fact that we do not account for the stopping time of a train at each station on the way, neither for the waiting time between two different trains at a station. From figure 3.5, we see that the observed travel time between Lyon and Bordeaux is stable from 1980 to 2005. Then, a drop of a hundred of minutes occurs. It comes from the fact that the trip between the two cities was previously made through normal lines,

crossing France from the East to the West. In 2001, the high-speed railways between Lyon and Marseille opened (line towards the south), which has been found to be convenient for the Lyon-Bordeaux trip, where trains run on a higher distance, but at a higher speed. However, as showed in figure 3.9, our computation with the Dijkstra algorithm considers that it is faster to pass through Paris after the openings of HSR between Lyon and Paris in 1981 and 1983 and between Paris and Tours in 1990, as showed by the variation in the estimated travel time at these years. In 2001, we see a slight variation in the estimated travel time since our computation considers that it becomes faster to pass through the South high-speed railway of Lyon-Marseille. After the opening of Tours-Bordeaux HSR in 2017, the trip passing by Paris becomes once again the preferred route.

3.6 Descriptive statistics

To present our data, we restrict our travel time dataset to the unique main city for each of the 94 NUTS3 region in metropolitan France (excluding Corsica). The selection of the main cities is done using INSEE data on the municipalities' population size from 1980 to 2020. We compute the average population size for each municipality and select the most populated one for each NUTS3 region.

Table 3.6 presents the statistics of train travel time growth for the $8,472$ pairs of cities.[8] Travel time time growth is computed as a growth rate between each year at which a main HSR roll-out has occurred with respect to 1980, prior any HSR roll-out. The pairs fully connected by HSR have observed a maximum decrease of travel time about 50%. Looking at the average travel time growth between main cities, we see that it has decreased from 5% in 1983 to nearly 20% in 2017 with respect to 1980. The median has evolved similarly over the years. Hence, the high-speed railways network may have impacted the inter-regional interaction opportunities of a majority of regions, not only the main cities at one end or the other of each high-speed line.

year	min	1st qu.	median	mean	3rd qu.	max	N
1983	-0.458	-0.022	0	-0.040	0	0	8,742
1990	-0.458	-0.142	0	-0.071	0	0	8,742
1994	-0.484	-0.163	0	-0.084	0	0	8,742
2001	-0.504	-0.200	-0.050	-0.110	0	0	8,742
2007	-0.526	-0.246	-0.113	-0.136	0	0	8,742
2017	-0.548	-0.316	-0.196	-0.188	0	0	8,742

Table 3.6: Train travel growth with respect to 1980

Table 3.7 shows the statistics on the HSR treatment intensity within each pair of NUTS3 regions. The HSR intensity is computed as the ratio of the distance on high-speed railways on the total rail distance, with rail distance approximated by the geodesic distance between every station in the path between to main cities. The first major wave of HSR roll-out has occurred in 1983 between Paris and Lyon. The pair is treated about a 100%, since the whole way is a HSR. Among all the pairs, 25% count between 10% to 100% rail distance treated by an HSR.

[8] We count 94 NUTS3 regions in metropolitan France excluding Corsica, which leads to $94 \times (94-1) = 8,472$ pairs of regions.

year	min	1st qu.	median	mean	3rd qu.	max	N
1983	0	0	0	0.121	0.096	1	8,742
1990	0	0	0	0.204	0.436	1	8,742
1994	0	0	0	0.238	0.499	1	8,742
2001	0	0	0.246	0.302	0.594	1	8,742
2007	0	0	0.342	0.357	0.662	1	8,742
2017	0	0	0.556	0.482	0.764	1	8,742

HSR intensity is computed as the ratio of the distance on high-speed railways on the total rail distance. Rail distance is approximated by the geodesic distance between every station in the path between to main cities.

Table 3.7: HSR treatment intensity

The second major wave occurs in 1990 with Paris connection to Le Mans and Tours. Among all the pairs, 25% count between 20% to 100% rail distance treated by an HSR. As time goes by, HSR covers more and more space within the country. Finally, about 50% of the pairs have at least 50% of the rail distance treated by HSR in 2017.

As showed in figure 3.1, the network has a star shape with France's capital city, Paris, at the center of the network and the other major cities at each end of the lines. Over time, the high-speed rail network has obviously reinforced Paris' centrality relative to train transportation network. However, the centrality of each municipality connected to railroads may also have changed substantially as discussed above. We define the municipalities' train transportation network centrality as the harmonic centrality index (Marchiori and Latora, 2000), which measures the importance of a node - a city - in a network, proportional to the sum of the inverse of its distances - travel time - from other nodes. The index is expressed as follows:

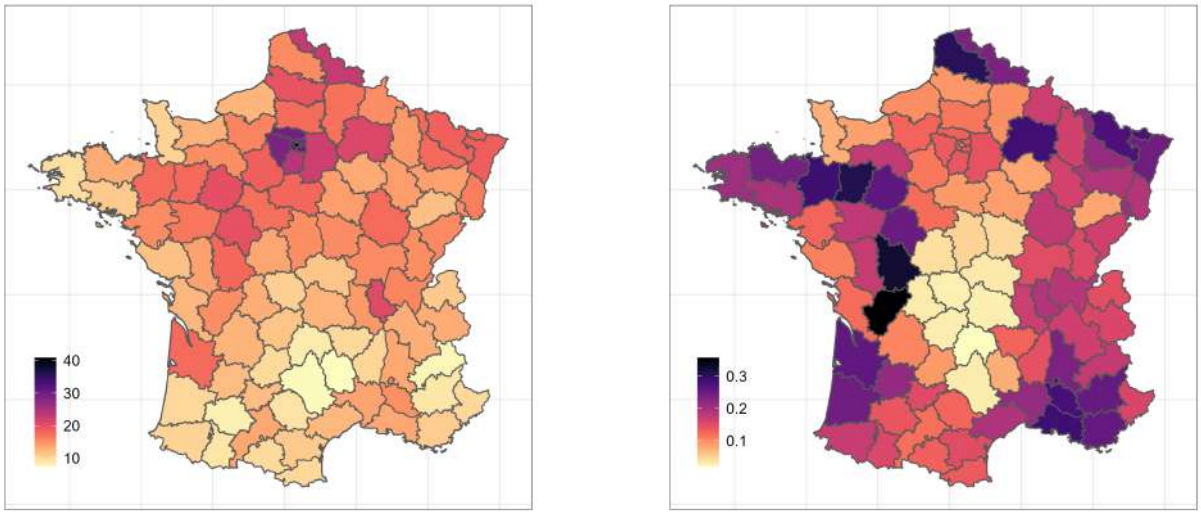
$$\text{centrality}_{it} = \sum_{j} \frac{1}{\text{travel time}_{ijt}} \tag{3.3}$$

with i and j the municipalities, t the year.

Figure 3.11a shows the 2020's centrality index values for the most populated city in each of the NUTS3 regions in France. The darker the color, the more central the region. Paris, with a centrality index about 41, and the regions around are shown to be substantially more central than the rest of France. Then, Lille is found to be in the 8^{th} position among the more central places, with a centrality index about 23. Reims, Lyon, Le Mans and Tours are following, with centrality index about 20. To put these results in a nutshell, cities connected by the HSR, or with a good connection to those cities, are the most central in 2020.

While the index of centrality is sensitive to the amount of stations present in the dataset, the log change in the centrality index is not. Figure 3.11b shows the log change in the centrality index between 1980 and 2020. The darker the color, the greater the change. The change in railroads centrality ranges from 0% to 36%, with lower values for regions located at the center of France and the highest values for regions directly connected to the HSR. Centrality of Paris and the surrounding regions has only increased by 12%, which is lower than the median in the sample.

While Paris may have been impacted the most from the HSR openings, its centrality index is not found to have changed considerably compared to other main cities. For example, main cities in Britain (the far west region on the map) have been increasingly central in the railway network, by more than 20%, while only Rennes is connected to Paris by an HSR. This example



(a) Centrality in 2020

(b) Centrality - log change (1980-2020)

Figure 3.11: Main cities' centrality relative to train transportation network

is to show that not only cities directly connected to the HSR may have been impacted, but since cities are all interlinked through the rail network, they all have experienced a shock in their accessibility, more or less intense.

3.7 Conclusion

The present paper describes a novel database of train travel time in France from 1980 to 2020 and the methodology to compute it. We use contemporaneous intercity trains' schedule datasets from SNCF. This data enable us to compute the contemporaneous travel time between every pair of municipalities in France using Dijkstra algorithm (Dijkstra, 1959). To compute past values on travel time, we identify non-stop stations pairs treated by high-speed railways (HSR) and the year of their roll-out, and make the assumption that prior to the HSR roll-out, trains were running at a normal speed. Then, we are able to compute the travel time between each pair of municipalities for each year going backwards in time. Comparing our estimations of travel time with a subsample of city-pairs from SNCF, we find that our estimations reproduce 95% of the total variation in the observed travel time. Descriptive statistics show that the high-speed network not only has impacted the cities directly connected to its rails, but also the whole set of French regions.

Chapter 4

Redefining Commuting: High-Speed Railways and Workers' Mobility in France

Abstract

This paper explores the influence of the French high-speed rail (HSR) network on workers' decisions related to workplace and residence. Utilizing data from the DADS Panel tout salarié and train travel time data from Gambuli and Stipanicic (2023), I employ a gravity model to measure the impact of reduced travel time on commuting adjustments between French departments (NUTS3 regions). Recognizing that HSR-affected department pairs still display long travel time for daily commuting, the analysis integrates considerations for telecommuting opportunities made feasible by internet connectivity. Stratifying the study by socio-professional categories accounts for varying telecommuting possibilities among worker groups. The findings indicate a growing tendency among workers, including executives, employees, and blue-collar workers, to choose separate and distant locations for residences and workplaces. This shift is attributed to both improved territorial connectivity through HSR and enhanced internet access. The combination of these factors offers both convenience and cost reduction for commuting, acting as mutually reinforcing elements.

Keywords: High-Speed Railways, Commuting
JEL Classification: O18, R40

4.1 Introduction

The advent of high-speed railways (HSR) has significantly enhanced regional accessibility by dramatically reducing travel times between cities and regions. Presently, as 55% of the workforce resides within an hour to an HSR station, these networks possess significant potential to influence commuting patterns.[1] Major cities that were once distant are now conveniently reachable within a reasonable commute time, typically ranging from 50 minutes to 2 hours. Given that HSR is designed for passenger transport, it can profoundly influence individuals' choices regarding their place of residence and employment, providing a convenient means of commuting.

Urban studies have extensively examined the impact of transport accessibility improvements on local commuting, labor markets, and city structures (Haas and Osland, 2014; Duranton and Puga, 2020). The literature typically depends on models of either monocentric or metropolitan urban structures, where individuals predominantly work in the city center, where most jobs are concentrated, and live in areas gradually extending towards the periphery, where housing tends to be more affordable (Alonso, 1964; de Palma et al., 2007; Mayer and Trevien, 2017).[2] However, the introduction of HSR introduces a novel dimension with implications at larger scales, enabling long-distance mobility for daily commuting between distant urban centers.

This paper recognizes the pivotal role of HSR in fostering inter-regional commuting, a phenomenon that has been less common compared to urban commuting up to this point. Adjustments in commuting flows may arise as individuals, by noting the new and lower commuting costs on certain region-pairs affected by the implementation of an HSR, can decide to change their residence, workplace, or both. They do so in order to choose the option of residence-work locations that maximizes their well-being, or utility. Hence, individuals might choose to live and work in distant locations if the utility derived from residing and working in those places surpasses the commuting costs from home to work, and especially if it provides the highest utility among all potential alternatives. While extensive research has focused on migration involving shifts in both residence and workplace from one place to another, this new type of migration where residence and workplace are apart, is less-explored in the literarure.

This research aims to contribute the the literature by qauntifying the effect of the introduction of HSR on workers' commuting behaviors and evaluate its potential impact on local economies. The significance of this aspect has spurred research initiatives in Spain, Germany, and China (Guirao et al., 2018; Heuermann and Schmieder, 2019; Wang et al., 2019; Feng et al., 2023). The present paper aims to address three key questions, unraveling how recent infrastructure improvements redefine commuting patterns in France.

Has the high-speed rail network shaped workers' commuting behavior in France? This study aims to examine the impact of improved connectivity on the spatial dyadic allocation of workers in terms of their residence and workplace. In particular, it identifies the causal impact of travel time reduction due to high-speed railways roll-out on the amount of commuters between NUTS3 region-pairs. Causality is identified using a gravity model with three ways

[1] Based on the population of NUTS3 regions and the train travel time between the main cities of NUTS3 regions.

[2] The first significant observation of interregional commuting was made during the reunification of Germany, after which residents of the East now had access to more attractive jobs and wages in the West while avoiding the often prohibitive costs associated with moving to areas where housing was more expensive (Burda and Hunt, 2001; Niebuhr et al., 2012; Ahlfeldt et al., 2015).

fixed effects.

The data show that high-speed rail network specifically affects travel time for long-distance region pairs. Commuters adjusting their residence-workplace locations over long distances due to HSR may experience significant commuting time despite the high speed of their commuting. Thus, their commuting time might not have decreased or, in some cases, may have increased. Telework, or remote work, is a component that can further enhance the affordability of long-distance commuting by alleviating parts of the commuting costs. It entails working from a location separate from the traditional workplace, often at home, using technology and internet access to fulfill their job responsibilities and stay connected with colleagues.

Telework has surged in popularity, driven by recent technological advancements, with a significant boost following the COVID-19 pandemic when telework became mandatory for a substantial portion of the population. Before this period, Hallépée and Mauroux (2019) reports that 11% of executives engaged in telework activities at least one day per week in France in 2017, against 3.2% of intermediate professions, 1.4% of employees, and 0.2% of blue-collar workers. This aligns with the conclusions drawn by Dingel and Neiman (2020), where they investigate the feasibility of remote work in the United States. Their research underscores that high-income occupations are generally more conducive to remote work when contrasted with low-income jobs.

An increasing body of research has shifted its focus towards understanding the effects of telework on commuting patterns, population density, and labor market outcomes (Gokan et al., 2022; Duranton and Handbury, 2023; Monte et al., 2023).[3] Telework is found to weaken agglomeration economies, playing as a centrifugal force. It aligns with the observed trend that telecommuting often leads to increased commuting distances (Nilles, 1991; de Vos et al., 2019), as well as suburbanization (De Fraja et al., 2021; Gokan et al., 2022; Liu and Su, 2023; Schulz et al., 2023).[4] This typically arise since teleworkers seek additional space in their homes for work, prompting a migration of their residence to more affordable locations that can accommodate their space needs. In France as of 2017, Hallépée and Mauroux (2019) indicate that 9% of the individuals residing more than 50 kilometers from their workplace engage in telework activities at least once a week, against to 1.8% of those within 5 kilometers. However, none of the studies investigated the complementary effect of telework and effective transportation mode (like high-speed rail).

The paper's second question seeks to address the following: *Has internet access a complementary impact to high-speed railways?* I investigate the associated impact implied by bilateral (residence and work) internet access, which has relevance to telework activities. To explore this, I construct an index at the NUTS3-region level, considering the evolving geographic coverage of ADSL and fiber optic technologies, as well as the dynamic internet speeds for both these technologies. I modify the commuting gravity model by introducing dummy variables for travel time reduction and internet access after they occur. Additionally, I include an interaction term to explore their potential complementary effects on commuting flows, resembling

[3]Specifically, Gokan et al. (2022) investigate the influence of telecommuting on the spatial organization of workers based on their skill types. They extend their model to include commuting between urban centers, suggesting that the HSR network could facilitate such commuting.

[4]Contrary to the centrifugal force of telework, Camagni et al. (2023) propose a view where 4.0 technologies serve as centripetal forces, reinforcing urban efficiency and countering potential diseconomies associated with large-scale agglomeration, showing that urban advantages persist even in the face of technological advancements in Italy. This paper suggests that the availability of 4.0 technologies in places can play as an amenitie valorized by firms and workers.

a triple difference-in-differences model. Furthermore, I segment the analysis by occupation groups to assess whether workers more inclined toward telework show greater responsiveness to long-distance commuting and reductions in travel time between residences and workplaces.

The third question this paper aims to answer is the following: *What are the main factors influencing changes in commuting patterns?* In a future version of this paper, I explore the key factors influencing changes in commuting flows between pairs of residence and workplace. This analysis involves estimating elasticity coefficients to travel time using distinct samples of different groups of individuals: those with a consistent place of residence who undergo changes in their workplace, those with a fixed workplace making residential adjustments, and those who make changes to both their workplace and residence. Furthermore, this study delves into the transitions between urban and rural environments, considering both residential and workplace location decisions. An additional focus is placed on examining the transitions on wages of people changing their workplace, comparing the levels at destination with respect to the origin workplace.

Preliminary findings reveal a growing inclination among workers, including executives, employees, and blue-collar workers, to choose separate and distant locations for their residences and workplaces. According to the estimated model, this trend can be attributed to the enhanced connectivity of regions facilitated by both HSR and internet. On one hand, an average travel time reduction of about 12% is found to be associated to a 6% increase in commuting flows on average, ceteris paribus. In the model with the dummy variable for travel time reduction, the associated effect without internet access is estimated to be the same: after any travel time reduction, commuting flows are expected to increase by 6% on average, in the absence of internet access.

On the other hand, internet access serves as a convenient amenity for individuals, as shown by its positive effect on commuting flows across all distance ranges, as well as a complement to HSR, indicated by the positive and significant coefficient for the interaction between travel time reduction and internet access. In particular, any travel time reduction for a residence-workplace pair with internet access is expected to increase the amount of commuters by about 14%. The overal increase in commuters due to HSR and internet over long distance (defined as exceeding 100 kilometers) is estimated to be about 30%.

Bilateral internet access may influence commuting flows and play as a complement of HSR by facilitating telework, thereby reducing the frequency of travel between home and work. This contributes to an overall reduction in weekly or monthly commuting costs, resulting in increased commuting flows over long distances. Indeed, the results show that bilateral internet access at both home and work have a more pronounced impact on commuting flows over longer distances, aligning with existing literature indicating that teleworkers tend to reside farther from their workplaces (de Vos et al., 2019; Hallépée and Mauroux, 2019; Schulz et al., 2023). This effect is even stronger for pairs that experienced a decrease in travel time through the implementation of HSR in the way, as mentionned previously.

The evolving trends in inter-regional commuting can bear significant consequences for regional disparities. In contrast to conventional migrants who relocate both their residence and consumption location, inter-regional commuters earn their income at the workplace and spend it in their (distant) homes. If these commuters primarily reside in peripheral regions and commute to high-wage urban areas, facilitated by enhanced connectivity to labor markets through HSR, the implementation of HSR can play a role in fostering economic development. This is achieved by boosting consumption capacity in remote areas, which can help reducing

regional disparities.

The structure of the paper is outlined as follows: Section 4.2 introduces the data employed in the analysis. In Section 4.3, the identification strategy is detailed, covering the examination of the HSR impact on commuting flows in Section 4.3.1 and exploring the complementary effect between HSR and internet access in Section 4.3.2. Section 4.4 presents the findings. Finally, Section 4.5 offers a discussion and concludes the paper.

4.2 Data

4.2.1 Data Description

This analysis involves the utilization of three main panel datasets on inter-regional workers' commuting flows, travel time by train and internet access.

Workers' Commuting

The dataset is based on the latest version of Panel Tous Salariés (INSEE), providing insights into a representative worker sample from 1976 to 2019. From 1976 to 2001, the data include information on workers born in October of an even year, while starting 2002, the sample has been expanded to add workers born one of the following 16 days: January 2 and 5, April 1 to 4, July 1 to 4, or October 1 to 4. To maintain sample consistency throughout the study period from 1993 to 2019, I keep individuals born in October of even years. This selection criterion ensures that the same cohort of individuals is considered throughout the entire analysis, allowing for meaningful comparisons and accurate assessments over time. The sample is assumed to be representative of the entire population of workers, given the assumption that the day of birth is random and does not influence future decisions on location, occupation, or commuting.

Each entry in the dataset corresponds to an individual's employment history within a company for a specific year. The dataset includes codes for the firm (SIREN) and the establishment (SIRET), the type of work contract (temporary, permanent, full-time, part-time), socioprofessional category aggregated in 6 groups,[5] the sector of activity, the annual net revenue, working hours, the year of entry in the firm and in the panel data, etc.

Additionally, and most importantly, it includes information about the locations of workplace and residence. Notably, residential data becomes accessible from 1993 onwards, and since that year, location details are recorded at the municipality level. Before 1993, it is available at the department level, which corresponds to the NUTS3 region definition of Eurostat. This location information enables the observation of the typical commute an individual undertakes in their daily life between home and work.

Building upon the approach outlined by Combes et al. (2012b), I refine the sample by specifically targeting individuals aged between 20 and 59 years who are engaged in full-time employment within the private sector.[6] This selection criterion is motivated by the desire to ex-

[5]The socioprofessional categories includes (1) farmers, (2) craftsmen, shopkeepers, and business leaders, (4) intermediate professions, including elementary school teachers, intermediate professions in health and social work, intermediate administrative and commercial professions in companies, technicians, (5) employees, and (6) blue-collar workers.

[6]It is worth noting that a small percentage, approximately 4% to 6% each year, may work in multiple establishments. To ensure consistency, I retain information from the main establishment, which is determined based on the establishment that provides the highest net revenue for the worker. I exclude observations outside continental

amine commuting behaviors and adjustments among individuals who are generally regarded as making stable decisions regarding their long term life and work.

The sample comprises 16,351,626 observations of 1,788,143 unique workers over a span of 27 years, from 1993 to 2019. On average, workers appear in the data for approximately 9.6 years during this time period, within a minimum of 1 year and a maximum of 35 years.

From this dataset, I compute the count of workers commuting from residence region i to work region j for each year. Despite the existence of 95 NUTS3 regions in continental France, I consolidate Paris and its three contiguous regions — Hauts-de-Seine, Seine-Saint-Denis, and Val-de-Marne — into a single region, collectively referred to as Paris' inner ring. This approach is justified by the extensive public transportation network interconnecting these regions, the widespread economic activity of Paris on these areas, and the significant volume of commuting between them. Given that the dataset spans 27 years in the panel, it results in a total of $N_i \times N_j \times N_t = 93 \times 93 \times 27 = 233,523$ observations.

To explore the heterogeneous effects of HSR and internet accessibility across various occupational groups, I extend the initial dataset. Specifically, I focused on the six aggregated groups defined by the occupational and socioeconomic classification *PCS ESE 2003*, with a particular emphasis on managers, intermediate professions, employees, and blue-collar workers. This expansion included the addition of four observations for each region-pair and year, wherein I computed the commuter count for each occupation. Consequently, the dataset expanded to include a total of $N_i \times N_j \times N_t \times N_o = 93 \times 93 \times 27 \times 4 = 934,092$ observations.

High-Speed Railways

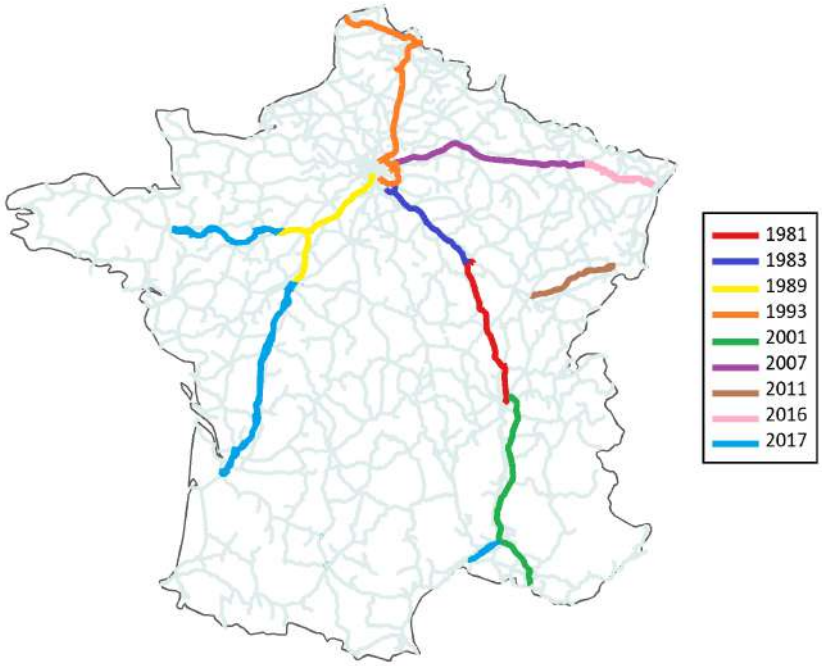
The second dataset focuses on the evolution of train travel time between municipalities in France from 1980 to 2020. This timeframe encompasses the implementation of high-speed rail (HSR) networks, starting with the initial HSR roll-out between Paris and Lyon in 1981-83, and concluding with the latest expansions in 2017, which include the lines connecting Le Mans to Rennes and Tours to Bordeaux.

As information regarding workers' residences is only accessible from 1993 onward, the analysis will reflect reductions in travel time starting from 2001. Notable instances include the introduction of the *Méditerranée* line from Lyon to Marseille in 2001, the extension from Paris to Strasbourg in 2007 with a final extension in 2016, the Dijon-to-Belfort connection in 2011, and three additional lines in 2017—linking Le Mans to Rennes, Tours to Bordeaux, and Montpellier to the *Méditerranée* line. Figure 4.1 illustrates the high-speed rail network, specifying the year of each line's opening.

Internet Access

Both geographical access to the internet and internet speed are crucial factors for internet accessibility. Internet connection speeds have evolved significantly, starting being measured in bykilobytes per second, advancing to megabytes with ADSL, and ultimately reaching gigabytes with fiber-optic technology. This evolution has greatly improved consumers' usage experience in terms of comfort and efficiency. The present paper computes a measure of internet access using three sources: panel data on ADSL and fiber-optic broadband geographical coverage at the municipality level, as well as the internet connection speed evolution over time for both technologies.

France and eliminate any entries lacking information on firms' entry year, residence, and workplace.



Notes: The map presents the high-speed rail network deployment in France. In 1981, the first segment of the LGV Sud-Est line, connecting Lyon to Paris, opened between Saint Florentin and Lyon. The second part, reaching Paris, opened in 1983. The LGV Atlantique saw its first branch, connecting Paris to Le Mans, open in 1989, followed by the second branch to Monts in 1990. Notable openings in subsequent years include the LGV Nord from Paris to Lille (1993), the LGV Rhône-Alpes from Lyon to Valence (1994), and the LGV Est from Paris to Baudrecourt (2007). The line towards Spain, Perpignan-Figuières, opened in 2010, and the LGV Sud Atlantique, extending the LGV Atlantique to Bordeaux, opened in 2017. Additionally, the LGV Bretagne-Pays de la Loire (connecting Le Mans to Rennes) and the bypass of Nîmes and Montpellier on the LGV Méditerranée both opened in 2017.

Figure 4.1: Fourty Years of High-Speed Railways Deployment (1981-2017)

First, Malgouyres et al. (2021) provides data on ADSL broadband access at the municipality level in France, spanning from 1995 to 2007. They computed a continuous measure of ADSL broadband access, with values ranging from 0 to 1, to capture the geographical coverage expansion of french municipalities across time. This measure has been computed using manual records of ADSL upgrade dates in telephone exchanges owned by the incumbent operator France Télécom (later renamed Orange) into which subscribers' telephone lines end at home or office. The data has further been completed with data from the French Regulatory Authority for Electronic Communications, Posts, and Press Distribution (ARCEP).[7]

Second, ARCEP supplies data on the geographical coverage of fiber-optic (FttH) infrastructure within French muncipalities from 2017 to 2023.[8] The FttH coverage rate of a municipality is a measure of the proportion of housing units or professional premises that can be connected to one or more FttH networks. This estimation involves the comparison of the total number of FttH lines deployed as reported by operators with an assessment of the overall number of premises within the municipality. The premises count is evaluated as the sum of residential housing and the count of business establishments with one or more employees, as determined by data published by INSEE.

Finally, I compile a dataset on the maximum internet connection speeds for the years 1993 to 2015, drawing information from an article published by GWS Media (2021), a digital media

[7] Autorité de régulation des communications électroniques, des postes et de la distribution de la presse

[8] Check French government website to get the data on the fixed broadband market deployment (ARCEP).

expert in the United Kingdom. The source provides a timeline of internet speeds during this period. According to Orange, the primary operator in France, current upload and download speeds can reach up to 10 Gbps. Leveraging this information and employing linear interpolation for years with missing information, I compile a dataset of internet speed records by year and technology, i.e. ADLS or fiber-optic.

4.2.2 Descriptive Statistics

This section presents the descriptive statistics.

Workers' Commuting Flows

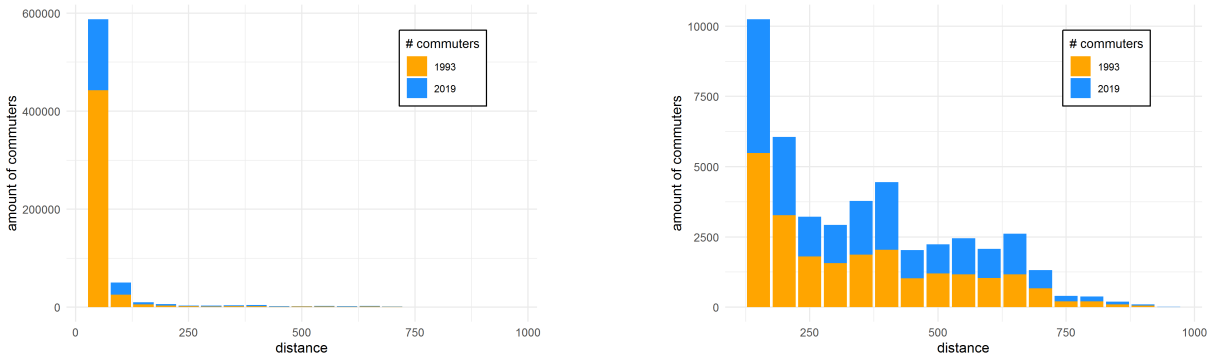
	Mean	SD	Min	Max
Residence Region				
1993				
Workers	5720.7	8049.5	344	70733
Out-commuters	1001.4	1669.5	51	8490
Share of out-commuters	16.8	9.9	6.3	61.3
Out-commuters $>$ 100km	285.9	203.3	30	976
Share of out-commuters $>$ 100km	6.4	2.4	1.1	14.2
2019				
Workers	7965.7	10945.3	611	95320
Out-commuters	1751.8	2477.4	82	13459
Share of out-commuters	22.3	11.1	8.9	67.7
Out-commuters $>$ 100km	555.1	475.0	37	2451
Share of out-commuters $>$100km	8.2	2.7	2.4	16.7
Work Region				
1993				
Workers	5720.9	10711.7	321	99740
In-commuters	1001.4	3769.5	28	35718
Share of in-commuters	12.0	6.5	4.7	40.6
In-commuters $>$ 100km	285.9	832.0	9	7904
Share of in-commuters $>$ 100km	4.1	1.7	1.5	13.0
2019				
Workers	7965.7	14329.3	627	131673
In-commuters	1751.8	5314.9	98	49812
Share of in-commuters	17.4	8.24	6.4	50.6
In-commuters $>$ 100km	555.1	1473.9	21	13764
Share of in-commuters $>$ 100km	5.8	2.6	1.9	18.3

Table 4.1: Descriptive Statistics of Regions

Inter-regional commuting is not a marginal phenomenon, as evident from the descriptive statistics in Table 4.1. In 1993, the share of out-commuters, defined as workers traveling to a region different from their residence, ranged from 6.3% to 61.3%, with an average of 16.8%. In 2019, this share ranged from 8.9% to 67.7%, with an average of 22.3%. Inter-regional commuting

is substantial and has increased over time. At longer distances, the share is still notable but comparatively lower. In 1993, an average of 6.4% commuted over distances exceeding 100 kilometers. This figure increased to 8.2% in 2019.The statistics in Table 4.1 reveal substantial standard errors, indicating considerable heterogeneity in terms of resident population and out-commuting flows.

Figure 4.2 illustrates the total number of commuters between residence and work regions categorized by distance thresholds (50 kilometers bandwidth), in 1993 and 2019. Notably, 2019's values consistently surpass those observed in 1993 across all distance ranges.



(a) Overall (b) distance $\geq$ 100km

Notes: This figure presents commuter counts between residence and work regions across distance thresholds (50 kilometers bandwidth). The orange bars denote values from 1993, the inaugural year of the panel, while the blue bars represent values from 2019, its final year. The left side depicts the overall distribution, while the right side zooms in on distances exceeding 100 kilometers. Notably, a substantial concentration of commuters occurs within 50 kilometers, overshadowing those at longer distances. However, the right side underscores a significant number of workers commuting over extended distances, with over 5,000 workers commuting between 100 and 150 kilometers in 1993, a number that doubled by 2019. Half the values are observed between 350 and 400 kilometers for both years. In 2019, values are consistently higher across all distance ranges compared to those observed in 1993.

Figure 4.2: Total amount of commuters by distance in 1993 and 2019

On the workplace side, Table 4.1 shows higher standard errors and notable variations between the minimum and maximum values concerning both the number of workers and in-commuters. In this context, in-commuters are individuals commuting to workplaces from their residences in different regions, viewed from the standpoint of the work region. These statistics unveil a pronounced concentration of employment, indicating that employment locations tend to be more clustered than residences. For instance, in 2019, the maximum number of out-commuters was about 13,459, while the maximum number of in-commuters was approximately 49,812 — almost four times higher. Notably, both statistics pertain to the Paris region.

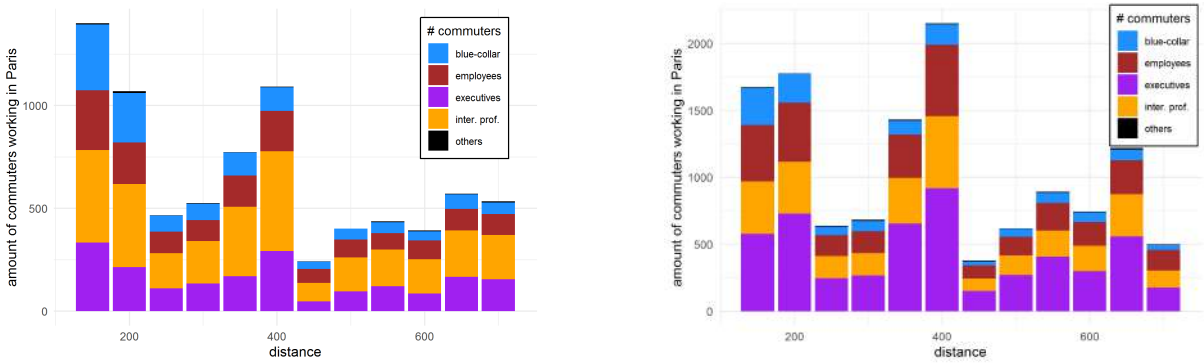
These findings underscore the critical need for attention to job opportunity accessibility. High levels of in-commuting highlight the concentration of employment in specific regions, emphasizing the importance of strategic interventions to promote balanced economic growth and enhance access to job opportunities. The implementation of HSR has the potential to play a pivotal role in achieving this objective by significantly decreasing travel times to major employment hubs. This is the focal point of investigation in the present paper.

Given that Paris is the region with the highest counts of both in-commuting (and out-commuting) flows and serves as a central hub within the HSR network, it can serve as an exemplary case for a more in-depth investigation into the number of commuters working in Paris based on distance thresholds from residence to workplace and the occupation composition.

In 1993, 8% of Parisian workers reside over 100 kilometers from their workplace, with 48% belonging to the intermediate profession group, 30% executives, 24% employees, and 19% blue-collar workers. In 2019, the corresponding figure increased to 10%, with 28% intermediate professionals, 48% executives, 29% employees, and 12% blue-collar workers.

Figure 4.3 shows that intermediate professionals were the primary cohort engaged in long-distance commuting in 1993, followed by executives. By 2019, there was a noteworthy reversal in this trend, with a substantial increase in long-distance commuting among executives.

Furthermore, in 2019, the number of workers commuting approximately 350-400 kilometers to work in Paris exceeded those residing at 100-150 kilometers, indicating a departure from the 1993 scenario and potentially signifying a noteworthy shift in commuting behaviors towards longer distances. For instance, within this distance range, the number of commuters residing in Lyon was 265 in 1993, and it rose to 915 in 2019. Similarly, there were 123 commuters living in Strasbourg in 1993, and this figure increased to 286 in 2019.



(a) 1993 (b) 2019

Notes: This figure illustrates commuter counts for individuals working in Paris based on the distance to their residence (with a 50-kilometer bandwidth threshold), with proportions indicated by socioprofessional category. The left side shows values for 1993, while the right side presents values for 2019. In 1993, intermediate professionals were the primary group in long-distance commuting, comprising approximately 48% of such commuters, followed by executives at around 30%. In 2019, executives became the majority, accounting for about 48% of long-distance commuters, while intermediate professions represented approximately 28%. Remarkably, in 2019, the count of workers commuting about 350-400 kilometers to work in Paris exceeds those residing at 100-150 kilometers from Paris, marking a notable shift from the scenario in 1993.

Figure 4.3: Total amount of commuters working in Paris in 1993 and 2019 by occupation (distance $\geq$ 100km)

Internet Access Measure

In the assessment of internet accessibility, I initially introduce an index at the municipal level. Subsequently, I aggregate this index at the NUTS3 region level. To achieve this, I rely on ADSL broadband access data from Malgouyres et al. (2021), fiber-optic broadband access data from ARCEP, and internet speed data from GWS Media (2021). At the municipal level, the index is computed as follows:

$$\begin{aligned}\text{Municipal Internet Access}_{mt} = \text{SpeedADSL}_t \times \text{CoverageADSL}_{mt} \times (1 - \text{CoverageFTTH}_{mt}) \\ + \text{SpeedFTTH}_t \times \text{CoverageFTTH}_{mt}\end{aligned} \tag{4.1}$$

where the weight factors, denoted as SpeedADSL and SpeedFTTH, correspond to the internet speeds of ADSL and FTTH at a given time t, respectively. The geographic coverage of these

technologies at time t in municipality m is represented by CoverageADSL and CoverageFTTH and take values between zero and one. Moreover, the ADSL coverage is appropriately weighted to account for the proportion of the population not covered by FTTH, which refers to the term $(1 - \text{CoverageFTTH}_{mt})$. By doing this, I prevent a significant increase in the index for areas where both ADSL and FTTH connections are available.

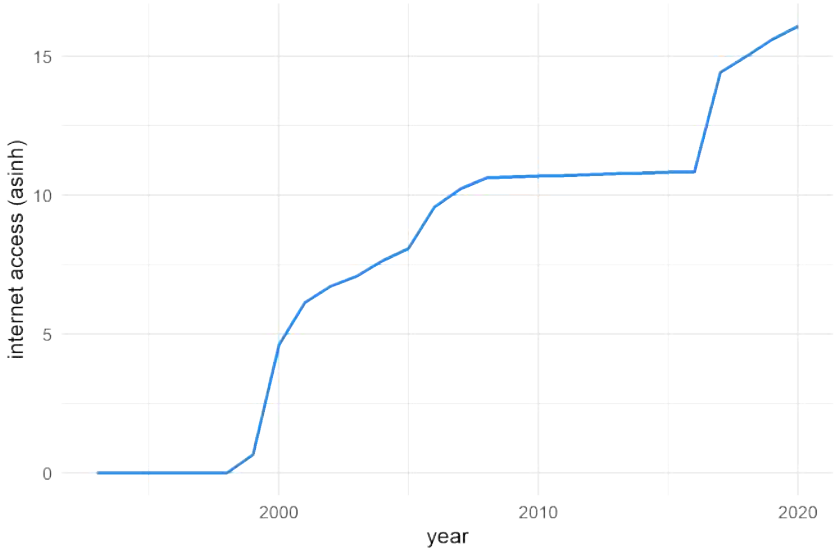


Figure 4.4: Average regional internet access over time

Subsequently, I aggregate the index at the regional level by calculating the weighted average of the municipal-level values using the following formula:

$$\text{Regional Internet Access}_{it} = \sum_{i} \text{Municipal Internet Access}_{m(i)t} \times \frac{\text{Population Size}_{m(i)t}}{\sum_i \text{Population Size}_{m(i)t}} \tag{4.2}$$

with municipality $m(i)$ located in region i. A weight is assigned to each municipality, determined by the proportion of the population in that specific municipality compared to the total population of the department. This approach ensures the regional representation of internet access aligns with its population distribution. The population data by municipality comes from the French National Institute of Statistics and Economic Studies (INSEE)[9]. Figure 4.4 displays the average values of regional internet access over time.

Finally, I define the bilateral internet access as follows, combining information of internet access at the residence i and workplace j:

$$\text{Bilateral Internet Access}_{ijt} = \text{Regional Internet Access}_{it} \times \text{Regional Internet Access}_{jt} \tag{4.3}$$

This measure, along with its binary counterpart that I refer to as $Internet_{ijt}$ (equal to one when *Bilateral Internet Access*$_{ijt}$ is strictly positive and zero otherwise), will be employed in the analysis. In evaluating the impact of internet connection, robustness checks will be conducted with different comparison values, such as the comparison to the median.[10]

[9]The population data by INSEE is available on the following website. It is available for the years 1990, 1999, and 2006 to 2020. For missing years in the data, I estimate the municipal population by linear interpolation using the last and the next observations available.

[10]When *Bilateral Internet Access*$_{ijt}$ surpasses the yearly median among region-pairs, the dummy variable

Final dataset

	Mean	SD	Min	Max
Dependant Variable				
1993				
Commuters	62.9	901.8	0	64022
Commuters $>$ 100km	3.3	14.7	0	443
Commuters $>$ 100km - Both in HSR	15.6	45.9	0	443
Commuters $>$ 100km - One in HSR	5.4	19.5	0	334
Commuters $>$ 100km - None in HSR	1.7	5.7	0	166
2019				
Commuters	87.5	1180.0	0	81861
Commuters $>$ 100km	6.5	29.1	0	915
Commuters $>$ 100km - Both in HSR	34.3	106.1	0	915
Commuters $>$ 100km - One in HSR	10.1	34.1	0	596
Commuters $>$ 100km - None in HSR	3.4	11.0	0	256
Communiting Costs				
Distance	391.2	188.5	31.5	988.2
1993				
Travel time	283.3	126.1	10.2	665.0
$\lvert\Delta^{\%}_{1980,1993}$travel time$\rvert$	7.3	10.8	0	46.3
Bilateral internet access	0	0	0	0
2019				
Travel time	242.1	100.6	10.2	547.1
$\lvert\Delta^{\%}_{1980,2019}$travel time$\rvert$	18.6	15.0	0	54.8
$\lvert\Delta^{\%}_{1993,2019}$travel time$\rvert$	12.2	11.9	0	53.9
Bilateral internet access	30.2	0.8	27.3	32.3

Table 4.2: Descriptive Statistics of Residence-Worplace Pairs

Table 4.2 presents the region-pair level statistics for both dependent and independent variables in the analysis. The average number of commuters per residence-workplace relation was approximately 63 in 1993 and increased to 88 in 2017, encompassing both intra-regional and inter-regional commuters. Specifically, for distances exceeding 100 kilometers between residence and workplace, the average was 3.3 in 1993 and rose to 6.5 in 2019. These figures are notably low, primarily due to the prevalence of zero values, with 44% of pairs exhibiting zero values in 1993 and 33% in 2019.[11] Additionally, the statistics indicate that commuting flows are higher for region-pairs where both regions have an HSR station by the end of the panel, compared to pairs where only one or none of the regions will be endowed with an HSR station.

Table 4.2 additionally presents statistics related to the proxies of commuting costs, including distance, travel time, and bilateral internet access for both 1993 and 2019. Over time, travel

$Internet_{ijt}$ is set to one; otherwise, it is set to zero. To prevent situations where region-pairs transition from having values of 1 to 0 due to fluctuations in median values, as a result of the median rising faster than a pair's bilateral connectivity, I maintain the dummy variable as 1 if it previously held a value of 1 at time t.

[11] Zero values pertain to region-pairs with an average distance greater than the average distance presented in Table 4.2, approximately 431 kilometers in 1993 and 442 kilometers in 2019.

time by train has decreased, attributed to the expansion of high-speed railways. Notably, there has been an average decrease of 12.2% in travel time from 1993 to 2019. It is noteworthy that before 1993, pairs had already experienced a 7.3% average decrease in travel time. This information will be considered in the identification strategy. Conversely, bilateral internet access has witnessed an increase, linked to the geographical expansion of ADSL since the early 2000s and fiber-optic from 2017, coupled with enhancements in their speed.

The subsequent section delineates the identification strategy, leveraging the aforementioned data, to establish a connection between commuting flows between residence and workplace, and commuting costs.

4.3 Identification Strategy

4.3.1 Adjustments in Aggregated Commuting Patterns

Baseline

Building on the established literature on commuting, I employ a gravity model to depict the connection between travel time and the volume of commuters. In this model, C_{ijt} represents the count of workers commuting from their residence in region i to their workplace in region j at year t. The gravity model takes the form:

$$C_{ijt} = R_{it}^{\alpha} W_{jt}^{\beta} T_{ijt}^{\gamma} \tag{4.4}$$

where R_{it} represents the time-varying overall probability to live in region i, influenced by its characteristics, such as the total amount of inhabitants, the regional amenities, housing prices, all of these variables impacting the average utility derived from living in region i; W_{jt} represents the time-varying overall probability to work in region j, influenced by its characteristics, such as the total amount of enterprises and workers, contributing to the average wage paid in workplace, thereby influencing the average utility derived from working in region j; T_{ijt} represents de commuting costs between the residence and the workplace, that are proxied bu the travel time by rail in this analysis. Commuting flows C_{ijt} are the result of the cumulative choices made by all workers regarding their residence and workplace locations (Ahlfeldt et al., 2015).

Equation 4.4 can be estimated using Pseudo-Poisson Maximum Likelihood model, suited for count dependant variables, which also include zero values and is efficient under heteroskedasticity (Silva and Tenreyro, 2006). Taking logarithms, putting it to the exponantial and adding an error term yields

$$C_{ijt} = \exp\Big[\alpha_{it} + \beta_{jt} + \gamma \log(T_{ijt})\Big]\eta_{ijt} \tag{4.5}$$

where α_{it} and β_{jt} are respectively the residence and workplace fixed effects, controlling for all possible regions' time varying characteristics R_{it} and W_{jt} previously mentionned, which help reducing the concern of omitted variable bias. Therefore, the estimation offers an effect of travel time while holding constant all potential time-varying characteristics at the regional level.

Fixed effects α_{it} and β_{jt} encompass information regarding the overall probability of workers residing and working in specific regions, taking into account factors such as amenities at

the residence, wages at the workplace, and the general connectivity of these places to others. When individuals evaluate all possible residence-workplace combinations, they select the option that maximizes their utility, considering all alternatives. As a result, these probabilities also reflect the likelihood of residing and working in these regions relative to all other potential locations, commonly referred to as *multilateral resistance terms*.

Since the objective is to provide the estimation of the effect of travel time reduction within pairs of residence-workplace, I incorporate pair fixed effects denoted as ρ_{ij}, which controls for all time unvarying unobservables for each pair, such as distance. Therefore, identification of the causal relationship is only possible by the changes in travel time between regions i and j.

$$C_{ijt} = \exp\Big[\rho_{ij} + \alpha_{it} + \beta_{jt} + \gamma \log(T_{ijt})\Big]\eta_{ijt} \tag{4.6}$$

Equation 4.6 allows us to properly estimate the elasticity coefficient γ, which represents the percentage effect of 1% travel time reduction between regions i and j on the number of workers commuting.

The Non-Random HSR implementation

While controlling for fixed effects greatly strengthens the reliability of the identification strategy by limiting omitted variables bias, changes in travel time induced by the implementation of HSR connection may be influenced by the growth in the amount of commuters, implying a reverse causality in the regression. This may be particularly true for pairs of cities directly connected by the high-speed rail network. However, the HSR network has had a significant impact on travel time between cities or regions, even if they are not directly connected to the network. It is especially the case for regions located by chance in the sourrounding of directly-connected regions. Those regions benefit from a good access to HSR station while they were not necessarility intended to directly benefit from the infrastructure.

To account for this endogeneity concern, I restrict the sample of region-pairs those that do not include an HSR station at the origin or/and destination. This way, I can examine the exogenous effect of travel time reduction, on commuting between regions that are indirectly impacted by the network due to their fortuitous proximity to cities where HSR stations have been implemented. Hence, coefficient γ is going to be the percentage effet of a 1% travel time reduction for pairs that do not have a direct HSR connection.

I also estimate regression 4.6 interacting travel time variable by including different dummies: (1) $\mathbb{1}(\text{Both in HSR}_{ijt})$ with value 1 if the pair ij has a direct HSR connection at year t, or an HSR station at both ends, zero otherwise; (2) $\mathbb{1}(\text{One in HSR}_{ijt})$, with value 1 if only one has an HSR station, zero otherwise; and (3) $\mathbb{1}(\text{None in HSR}_{ijt})$, with value 1 if none of the residence and work locations have an HSR station, zero otherwise. Thus, three coefficients are estimated for each of these groups. These results will be presented in the section of the analysis that explores heterogeneous effects of travel time on commuting flows.

The Case of Already Treated Pairs. Given that the information on residence is only available from 1993 onwards, while the high-speed rail (HSR) network was introduced in 1981 and expanded to multiple segments during that period, estimating the impact of travel time reduction on commuting between 1993 and 2019 would involve comparing the growth of commuting in newly treated pairs with already treated pairs. In particular, the pair fixed effects might entirely incorporate the effects of the already treated pairs if they do not experience an

additional decrease in travel time after 1993. This could introduce identification challenges and potentially lead to biased estimates, which I expect to be lower in magnitude than their true value since already treated pairs may experience an increase in their commuting flows.

To address this concern, I exclude pairs that were already treated, identifying them as those with HSR stations at both ends before 1993. To further test for robustness of the results, I additionally omit pairs that exhibited any decrease in travel time prior to 1993, even if not directly connected to the network. Will remain the pairs solely impacted by the HSR connections of Lyon-Marseille (2001), Paris-Baudrecourt (2007), Dijon-Belfort (2011), Baudrecourt-Strasbourg (2016), Le Mans-Rennes (2017) and Tours-Bordeaux (2017). This approach aims to address any potential issues associated with the differential treatment effect.

Heterogenous Effects

The paper delves deeper into the analysis by exploring the heterogeneous effects of travel time reduction across various distance thresholds. The underlying premise is that the high-speed rail network has made commuting feasible for region pairs that were previously characterized by unaffordable travel times. This effect is anticipated to be more pronounced for pairs separated by relatively long distances.

Additionally, I examine the heterogeneous impact of travel time reduction based on the presence of HSR stations in both, one, or neither of the regions. Even pairs without an HSR station can experience a decrease in travel time if an HSR route is part of the commuting path. This scenario arises when the worker takes a train on regular lines from the residence to the workplace origin, which then transitions to high-speed railways and normal lines upon reaching the destination station. Consider the case of individuals residing in Saint-Etienne and working in Toulon, or vice versa, facilitated by the *Méditerranée* high-speed line.

Heterogenous effects of travel time reduction on commuting flows are also investigated across different occupation groups: executives, intermediate professions, employees and blue-collar workers. Workers from diverse occupations may have distinct preferences, valuations of time, and levels of flexibility in their work arrangements. Executives, for instance, are likely to place a higher value on their time, making reductions in travel time more influential in their commuting decisions compared to employees or blue-collar workers. Additionally, variations in income and resources among occupational groups can influence their ability to adapt to changes in commuting costs, such as affording high-speed trains tickets or be able to relocate their residence easily.

4.3.2 High-Speed Railways, Internet Access and Telework

Complementary Effect of HSR and Internet Access

Considering that travel time between regions connected by high-speed rail remains substantial for daily commuting between residence and workplace, typically more than an hour, a pertinent question arises regarding the correlation between workers who maintain a significant distance between their home and workplace and their likelihood to participate in telework, or work partially from home. This hypothesis stems from the idea that individuals who opt for distant residence-work setups might exhibit a greater inclination towards flexible work solutions, like telecommuting, which allows them to mitigate the challenges associated with long-distance commuting.

In this preliminary analysis, I investigate the combined impact of travel time reduction and internet access, considering the latter's role in facilitating telework. Barrero et al. (2021) support this idea that the feasibility of telework is often contingent on the effective use of information and communication technologies. In particular, they demonstrate that enhanced internet access not only directly improves telework efficiency but also broadens the scope of teleworking, holding the worker's earnings and industry of employment constant.[12]

Internet access can also serves as an amenity for both residences and workplaces. Regions with internet access can become more attractive to households, offering access to various leisure activities and access to a range of consumption goods. Simultaneously, internet access becomes appealing to labor markets, for firms that rely on it for production and communication. The increasing demand for labor of certain firms, stemming from the need for internet-based production or communication, can attract workers from remote locations. Alternatively, it may compel individuals to offer their labor if opportunities closer to home are limited or attract them when heightened demand pushes wages upward.

I integrate those sets of information in a gravity equation, where the bilateral internet access at home and work plays as a factor reducing commuting costs and as an amenity in both residence and work locations, integrated in the fixed effects α_{it} and β_{jt}. The estimated model is written as follows:

$$C_{ijt} = \exp\left[\rho_{ij} + \alpha_{it} + \beta_{jt} + \gamma\,\text{HSR}_{ijt} + \delta\,\text{Internet}_{ijt} + \theta\,\text{HSR}_{ijt} \times \text{Internet}_{ijt}\right]\eta_{ijt} \tag{4.7}$$

where the variable HSR_{ijt} is defined as one after the region-pair experience a travel time reduction due to the implementation of an high-speed rail line, zero otherwise. The variable $Internet_{ijt}$ is defined as one after the region-pair has access to internet (for strictly positive values in the variable *Bilateral Internet Access*$_{ijt}$), zero otherwise. Additionally, α_{it}, β_{jt} are ρ_{ij} are the set of fixed effects needed for causal interpretation, as presented previously.

The coefficient γ represents the average percentage point effect of travel time reduction in the absence internet access, while the coefficient δ represents the average percentage point effect of internet access in the absence of an HSR connection. Lastly, a positive sign and significance of the coefficient θ would assess the complementary effect of internet access and travel time reduction. The combined impact of both technologies will be interpreted as $(\gamma + \delta + \theta) \times 100\%$ change in the number of commuters between residence and work regions.

Finally, replacing $Internet_{ijt}$ with *Bilateral Internet Access*$_{ijt}$ in the hyperbolic inverse function enables the interpretation of results in terms of the percentage change in internet access, similar to the logarithmic function, and accommodates zero values. This approach assesses the impact of time-varying internet access intensity.

Investigation by Workers' Occupation

Hallépée and Mauroux (2019) examine the proportion of workers engaged in regular telework, categorized by their socio-professional groups in France. The authors consider telework as "regular" when it takes place at least once a week, adhering to the definition provided in Article 2 of the 2002 European Framework Agreement on Telework. This framework defines telework as a work organization method that leverages information and telecommunications

[12] Barrero et al. (2021) draws on information sourced from the American Survey of Working Arrangements and Attitudes encompassing the period prior to and following the onset of the Covid-19 pandemic.

technologies, operating within the confines of a contract or employment agreement and enabling tasks typically performed on-site to be completed remotely on a consistent basis. They use data sourced from the Sumer[13] 2017 survey, which stand for questions on medical monitoring of employee exposure to occupational risks. According to their findings, 11.1% of executives practice telework at least on day per week, against 3.2% of intermediate professions, 1.4% of employees, and 0.2% of blue-collar workers.

Panel Tous Salariés (INSEE) data provide information on aggregated occupation group for each worker, which I use in order to compute the amount of commuters by region-pair and occupation for each year of the panel. In this part of the analysis, I examine the effect of internet access, coupled with travel time reduction between regions, on long-distance commuting, keeping in mind the propensity to telework by occupation in France provided by Hallépée and Mauroux (2019).

Similarly, I estimate the following equation:

$$C_{ijot} = \exp\Big[\rho_{ijo} + \alpha_{iot} + \beta_{jot} + \sum_{\text{occ}=1}^{4} \gamma_{\text{occ}}\, \text{HSR}_{ijt} \times \mathbb{1}(\text{occ} = o) + \sum_{\text{occ}=1}^{4} \delta_{\text{occ}}\, \text{Internet}_{ijt} \times \mathbb{1}(\text{occ} = o) + \sum_{\text{occ}=1}^{4} \theta_{\text{occ}}\, \text{HSR}_{ijt} \times \text{Internet}_{ijt} \times \mathbb{1}(\text{occ} = o)\Big]\eta_{ijot} \tag{4.8}$$

where occupation group are represented by the subscript o, with $o = \{$*executives, intermediate professions, employees, blue-collar workers*$\}$. Variables α_{iot} and β_{jot} are respectively the residence-occupation and workplace-occupation fixed effects, controlling for all possible regions' and occupations' time varying characteristics influencing the overall probability for people to respectively live and work in these regions according to their occupation. Variable ρ_{ijo} represent the set of fixed effects for all possible factors influencing commuting costs which do not vary over time. For each of the occupational group, the γ, δ and θ coefficients are estimated.

4.3.3 Adjustments Margins

Furthermore, I explore the different dimensions of adjustments, including the mobility of residence versus workplace and urban versus rural adjustments, as Heuermann and Schmieder (2019), as well as the evolution of wages, which go beyond Heuermann and Schmieder's work. To do so, as they do, I segment the sample of workers into distinct groups and conduct separate regressions for each group. This approach allows to investigate the specific dynamics and factors that influence each adjustment margin following travel time reduction.

Residence versus Workplace Mobility. For this part of the analysis, the dependant variable is computed as the count of commuters that travel to destination j - workplace - conditional on the fact that they are not moving their origin location i - residence. On another hand, I consider the amount of commuters that travel from origin i - residence - conditional on the fact that

[13] Surveillance médicale des expositions des salariés aux risques professionnels

they are not moving their destination location j - workplace. Finally, I consider the sample of workers that adjust both their residence and work locations. Doing so give insights on which margin between the employment or residential drives results found estimating equation 4.6. In the German context, Heuermann and Schmieder (2019) find that commuting pattern adjustments predominantly stem from changes in the workplace rather than the residence. This pattern may be attributed to the housing market being less flexible than the labor market, a distinction commonly identified in the literature Haas and Osland (2014). The present paper is going to verify this relation in the context of France.

Larger Versus Smaller Region. I categorize commuters based on the relative size of the origin and destination regions. This classification yield three distinct groups for origin-destination pairs ij: pairs where the residence region is larger than the work region, pairs where the residence region is smaller than the work region, and pairs where the regions are of equal size. The dependant variables are computed as the count of commuters within each of these three groups, and differenciate between urban and rural areas according to Eurostat Classification of NUTS3 regions. While Guirao et al. (2018) provides evidence supporting an increase in commuting flows from small residence regions to large employment regions due to improved connectivity in Spain, Heuermann and Schmieder (2019) reports the opposite trend in Germany: people tend to commute more from large residence regions to smaller workplace regions. This unexpected result is influenced by the location of headquarters or main production sites of several large and highly-productive firms situated in smaller cities rather than urban centers, including Audi, Volkswagen, and Siemens. The present paper is going to investigate the direction of such patterns in the case of France.

Wage Evolution. Using the sample of workers who change their workplace, I am going to identify groups according to the evolution of their wage at destination - new workplace. I count the amount of commuters for two different groups: (1) individuals who experienced an upgrade in wage at the new workplace with respect to their previous workplace, and (2) those who experienced a downgrade or no change. The literature identifies the first channel as the most significant factor driving migration, a conclusion supported by the review conducted by Haas and Osland (2014). The present paper is going to investigate this point.

4.4 Results

4.4.1 Baseline

Evidence of localized commuting

Before delving into the causal analysis, I provide compelling evidence that workers' commuting behavior is localized and negatively correlated with distance. In Table 4.3, I present the estimated results of equation 4.5, utilizing distance and travel time variables. The first three models incorporate region-year fixed effects for both residence and workplace, evaluating the impact of distance and travel time by train between region-pair while holding region-specific time-varying characteristics constant. The fourth model further includes region-pair fixed effects, enabling a causal assessment of the travel time reduction effect within pairs.

As expected, the results indicate a negative correlation between distance and commuting,

with a coefficient of approximately -3.43 (see column 1). This suggests that if region i is 1% closer to region j than to another region k, the expected number of commuters living in region i and working in region j is 3.48% higher compared to those heading toward region k. The coefficient magnitude greater than 1 suggests a substantial impact of distance on commuting flows, emphasizing that commuting costs increase substantially with distance.

The coefficient for travel time remains consistent, about -2.29 (refer to column 2). Upon incorporating both distance and travel time, the analysis indicates that, for a given distance between regions i and j and between i and k, and with a one-percentage-point difference in travel time, the difference in commuting flows is estimated to be around 1.40 percentage points (see column 3). This finding underscores the potentially substantial impact of reducing travel time, which could outweigh the influence of distance in shaping commuting costs. Faster commuting over distances may lead to noticeable shifts in commuting patterns and the spatial distribution of workers.

	Model 1	Model 2	Model 3	Model 4
log(distance$_{ij}$)	-3.43^{***}		-1.28^{***}	
	(0.04)		(0.11)	
log(travel time$_{ijt}$)		-2.29^{***}	-1.40^{***}	-0.49^{***}
		(0.02)	(0.08)	(0.03)
Fixed Effects				
ρ_{ij}	No	No	No	Yes
α_{it} and β_{jt}	Yes	Yes	Yes	Yes
N	223587	223587	223587	223587
Pseudo R^2	0.96	0.97	0.98	0.99

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. The first three models incorporate region-year fixed effects for both residence and workplace, evaluating the impact of distance and travel time by train while holding region-specific time-varying characteristics constant. The fourth model further includes region-pair fixed effects, enabling an assessment of the travel time reduction effect within pairs.

Table 4.3: # commuters, distance and travel time

In the baseline regression presented in equation 4.6 and estimated in column 4, I account for both region-time and region-pair specific effects. By introducing pair fixed effects, I can estimate the average impact of travel time reduction on the number of commuters within region pairs. These fixed effects allows a comparison of these trends between pairs of regions impacted by the high-speed rail network and pairs that are not impacted, effectively serving as a control group for the analysis. Additionnaly, as I estimate an elasticity within a log-linear relationship, the model also enables a comparison between pairs experiencing varying degrees of travel time reduction.

Results demonstrate that a 12% reduction in travel time between regions—representing the average travel time reduction—while accounting for regional characteristics of residence and work, is associated with a 6% increase in commuter numbers on average (see column 4). For pairs characterized by a direct HSR connection (absence of a regular line in between) and devoid of any intermediary station, the average reduction in travel time of 30% is associated with a notable 15% increase in commuting flows.

	Model 1	Model 2	Model 3	Model 4	Model 5
asinh(internet access$_{ijt}$)	1.06^{***}	0.16^{***}	0.01	0.05^{***}	0.02^{***}
	(0.14)	(0.03)	(0.02)	(0.02)	(0.00)
log(distance$_{ij}$)		-3.43^{***}		-1.28^{***}	
		(0.04)		(0.11)	
log(travel time$_{ijt}$)			-2.29^{***}	-1.40^{***}	-0.48^{***}
			(0.02)	(0.08)	(0.03)
Fixed Effects					
ρ_{ij}	No	No	No	No	Yes
α_{it} and β_{jt}	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.26	0.96	0.97	0.98	0.99
N	223587	223587	223587	223587	223587

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. The first four models incorporate region-year fixed effects for both residence and workplace, evaluating the impact of distance and travel time by train while holding region-specific time-varying characteristics constant. The fifth model further includes region-pair fixed effects, enabling an assessment of the travel time reduction and internet access improvement effects within pairs.

Table 4.4: # commuters, distance, travel time and internet access

Internet Access: Regional Amenity and Telecommutability

Controlling for regional characteristics over time, column 1 of Table 4.4 provides evidence that heightened commuter flows are anticipated between residences and workplaces with improved internet connectivity, as indicated by an elasticity coefficient of 1.1. However, when controlling for distance alone, the coefficient decreases to 0.2 (column 2) and further drops to 0.05 when additionally accounting for travel time (column 4). As evidenced by the upward bias in the coefficient of internet access in column 2 compared to column 4, there exists a negative correlation between internet access and travel time. This suggests that higher travel times align with increased bilateral internet access, a trend reflective of France's geographic layout characterized by a polycentric system of cities. Wealthier and more developed regions, situated at significiant distances and travel times from each other, demonstrate enhanced internet connectivity in a given year.

Controlling for all potential time-invariant commuting costs through the utilization of region-pair fixed effects, we assess the impact of enhanced bilateral internet access within residence-workplace pairs. For every 1% increase in internet access, commuting flows are anticipated to rise by 0.02%. The travel time reduction effect is found to be 0.01 percentage points lower only when controlling for internet access.

Improved internet connectivity at both the residence and workplace is expected to boost commuting flows. However, travel time continues to play a crucial role in commuting decisions. Despite the growth of telecommuting, which was practiced by only around 3% of the population at least once per week before the COVID-19 pandemic, as evidenced by Hallépée and Mauroux (2019), 30% of those telecommuters worked from home three days or more per week. Even with the rise of telecommuting, workers are expected to maintain a regular commuting routine, highlighting the continued significance of travel time in the decision-making process of choosing a place to live and work.

Robustness

The last estimation in Table 4.4 (column 5) raises two notable concerns. Firstly, some region pairs had already experienced a decrease in travel time before the sample's starting date in 1993. For instance, pairs like Paris-Lyon, connected between 1981 and 1983, or Paris-Lille, connected in 1993, are examples of such cases, with an HSR station at both ends. As a result, these pairs may still be influenced by the past improvement in connectivity when observing the evolution of the number of commuters. If they do not experience further decreases in travel time after 1993, they become absorbed in the pair fixed effects and inadvertently serve as a control group. This could introduce bias into the estimator, potentially leading to a reduction in the magnitude of the coefficient. Pairs that encountered a reduction in travel time both before and after 1993 are additionally excluded from the sample for analogous reasons. An illustration of such pairs is Paris-Marseille.

To tackle this concern, I eliminate the pairs that have already undergone treatment from the sample in columns 1 and 2 of Table 4.5. Column 1 excludes pairs with HSR stations at both ends before 1993, denoted as *criteria 1*, while column 2 goes a step further by excluding all pairs that have experienced any decrease in travel time, referred to as *criteria 2*.

	(1)	(2)	(3)	(4)	(5)	(6)
log(travel time)	-0.49^{***}	-0.47^{***}	-0.53^{***}	-0.53^{***}	-0.69^{***}	-0.76^{***}
	(0.03)	(0.04)	(0.04)	(0.05)	(0.07)	(0.12)
asinh(internet access)	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}	-0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sample						
Exclude if already treated: criteria 1	Yes	Yes	Yes	Yes	Yes	Yes
Exclude if already treated: criteria 2	No	Yes	No	Yes	No	Yes
Exclude if i and j have HSR station	No	No	Yes	Yes	Yes	Yes
Exclude if i or j have HSR station	No	No	No	No	Yes	Yes
α_{it} and β_{jt} FE	Yes	Yes	Yes	Yes	Yes	Yes
ρ_{ij} FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.99	1.00	0.99	1.00	0.99	1.00
N obs	217,485	120,285	210,978	111,179	143,559	86,589

Table 4.5: Endogeneity concerns - sample selection

The second concern relates to pairs that witnessed the most substantial decrease in travel time, which is primarily driven by the presence of an HSR station. The location of an HSR station typically indicates that the region or city is a significant center of economic activity, attracting commuters from various locations. To control for this size effect which raise reverse causality concern, I further refine the sample. In column 3 and 4, I exclude pairs where both i and j have an HSR station. Then, in column 5 and 6, I exclude pairs where either i or j has an HSR station. This selection process retains only those pairs that experienced a decrease in travel time due to their fortunate proximity to directly connected regions, allowing for a more accurate analysis of the impact of travel time reduction on the evolution of commuting patterns.

The results provide reassurance as the significance of the elasticity coefficient remains even after the sample selection. Additionnally, the coefficient does not change when excluding already treated pairs (columns 1 and 2). Results indicate that the impact of travel time reduction on commuting behavior remains strong and consistent across different sets of region

pairs, including those that may not have been specifically designed to increase the number of commuters. This suggests that the influence of reduced travel time on commuting patterns extends beyond regions directly connected to the HSR and holds true across various contexts and locations, those that have indirectly benefited from the infrastructure.

Interestingly, we observe an higher magnitude in the effect of travel time reduction when excluding regions-pairs endowed with an HSR station. One possible explanation for this observation is that pairs not connected by the high-speed railway network generally exhibit fewer commuters, as evident from the descriptive statistics in Table 4.2. As a result, these pairs are inherently more sensitive to changes in the number of commuters when travel time is reduced. In contrast, pairs with a high level of commuting activity might show a more stable pattern in the face of variations in travel time.

4.4.2 Heterogeneity

Table 4.6 presents the results examining the heterogeneity of the travel time reduction effect. First, the analysis is stratified based on groups of regions according to their connectivity to the HSR network, identified by the presence of an HSR station in the region in 2019.[14] This categorization includes scenarios where both ends have an HSR station in the final year of the panel ($\mathbb{1}_{\text{Both in HSR}}$), only one end has an HSR station ($\mathbb{1}_{\text{One in HSR}}$), and no HSR station at both ends ($\mathbb{1}_{\text{None in HSR}}$), with the dummy variable taking the value of one or zero accordingly. Second, the analysis explores heterogeneity based on different distance thresholds.

The results indicate that the impact of travel time reduction is more pronounced for residence-workplace pairs of regions without an HSR station, confirming results found in the previous robustness section. In these pairs, there is a reduction in travel time as an HSR is present along the route, but the connection to the HSR is indirect and somewhat arbitrary due to their fortuous position and proximity to regions with an HSR station. For such pairs, the average reduction in travel time is approximately 9%, leading to an expected average increase of 6% in the number of commuters within the pair.

The impact of travel time reduction also varies based on distance, exhibiting greater effects at distance thresholds of $[100; 200]$, $[400; 500]$, and distances exceeding 600 kilometers. Notably, these distance thresholds correspond to typical distances between major French regions. For instance, typical examples of major city pairs within the 100 to 200 kilometers range are Paris-Lille, Paris-Calais, Paris-Reims, Reims-Metz, Tours-Angoulême, Marseille-Nîmes, and Lyon-Dijon. In the 400 to 500 kilometers range, we find pairs such as Paris-Bordeaux, Paris-Lorient, Lyon-Reims, Lyon-Strasbourg, and Lille-Angers.

This evidence supports the non-linear nature of the relationship between commuting flows and distance, due to distorted relationship between distance and travel time. Given that region characteristics are held constant in the econometric estimation, the higher elasticity coefficients associated with larger distances highlight differences in the initial volumes of commuting flows. Specifically, these coefficients suggest that initial commuting volumes were lower at larger distances.

Further heterogeneity evidence is provided by workers' occupation group among executives, intermediate professions, employees, and blue-collar workers. Results are displayed

[14] I use the year 2019 as the reference year for identifying regions with an HSR station. This choice is made to avoid pairs switching from one group to another, allowing for an assessment of the impact of travel time reduction on pairs that will be endowed (or not) by the end of the panel.

	Model 1	Model 2	Model 3
$\log(\text{travel time}_{ijt})$	-0.48^{***}		
	(0.03)		
$\text{asinh}(\text{internet access}_{ijt})$	0.02^{***}	0.02^{***}	0.02^{***}
	(0.00)	(0.00)	(0.00)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{\text{Both in HSR}}$		-0.45^{***}	
		(0.04)	
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{\text{One in HSR}}$		-0.48^{***}	
		(0.04)	
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{\text{None in HSR}}$		-0.67^{***}	
		(0.07)	
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{\text{distance} \leq 100}$			-0.45^{***}
			(0.07)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{100 < \text{distance} \leq 200}$			-0.57^{***}
			(0.07)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{200 < \text{distance} \leq 300}$			-0.46^{***}
			(0.09)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{300 < \text{distance} \leq 400}$			-0.40^{***}
			(0.07)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{400 < \text{distance} \leq 500}$			-0.51^{***}
			(0.08)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{500 < \text{distance} \leq 600}$			-0.43^{***}
			(0.06)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{600 < \text{distance} \leq 700}$			-0.60^{***}
			(0.07)
$\log(\text{travel time}_{ijt}) \times \mathbb{1}_{\text{distance} > 700}$			-0.76^{***}
			(0.12)
Fixed Effects			
ρ_{ij}	Yes	Yes	Yes
α_{it} and β_{jt}	Yes	Yes	Yes
Pseudo R^2	0.99	0.99	0.99
N	218295	218295	218295

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. All models incorporate region-year fixed effects for both residence and workplace and region-pair fixed effects, enabling an assessment of the travel time reduction and internet access improvement effects within pairs.

Table 4.6: Heterogeneity in the effect of travel time reduction on commuting flows

in Table 4.7. In the specification with the usual fixed effects ρ_{ij}, α_{it} and β_{jt}, executives are found to be the more responsive to travel time reduction for their commuting through their joint decision of residence-workplace location choice, followed by intermediate professions, employees and finally blue-collar workers.

Moreover, executives are those the most impacted by bilateral internet access, especially at long distances. They are also the one that may use more internet than the others. This ranking is the effect of travel time reduction and internet access improvement is in line with the telework propensity found by Hallépée and Mauroux (2019).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Executives						
log(travel time$_{ijt}$)	-0.68^{***}	-0.67^{***}	-0.45^{***}	-0.25^{***}	-0.25^{***}	-0.19^{***}
	(0.07)	(0.05)	(0.06)	(0.04)	(0.04)	(0.05)
asinh(internet access$_{ijt}$)		0.02^{***}	0.02^{**}		0.01^{**}	0.02^{**}
		(0.01)	(0.01)		(0.01)	(0.01)
asinh(internet access$_{ijt}$)$\times\mathbb{1}_{\text{dist}>100}$			0.06^{***}			0.00^{***}
			(0.01)			(0.00)
Intermediate Professions						
log(travel time)	-0.54^{***}	-0.48^{***}	-0.22^{***}	-0.07	-0.07	-0.19^{***}
	(0.05)	(0.05)	(0.06)	(0.04)	(0.04)	(0.04)
asinh(internet access$_{ijt}$)		0.01^{**}	0.00		0.00	0.00
		(0.00)	(0.00)		(0.00)	(0.00)
asinh(internet access$_{ijt}$)$\times\mathbb{1}_{\text{dist}>100}$			0.02^{***}			-0.00^{***}
			(0.00)			(0.00)
Employees						
log(travel time)	-0.49^{***}	-0.53^{***}	-0.26^{***}	-0.79^{***}	-0.78^{***}	-0.30^{***}
	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)	(0.05)
asinh(internet access$_{ijt}$)		0.03^{***}	0.02^{***}		0.05^{***}	0.03^{***}
		(0.00)	(0.00)		(0.01)	(0.00)
asinh(internet access$_{ijt}$)$\times\mathbb{1}_{\text{dist}>100}$			0.00			0.02^{***}
			(0.00)			(0.00)
Blue-Collar Workers						
log(travel time)	-0.37^{***}	-0.38^{***}	-0.08	-0.72^{***}	-0.72^{***}	-0.32^{***}
	(0.05)	(0.05)	(0.06)	(0.07)	(0.07)	(0.05)
asinh(internet access$_{ijt}$)		0.02^{***}	0.01^{***}		0.02^{***}	0.01^{***}
		(0.00)	(0.00)		(0.00)	(0.00)
asinh(internet access$_{ijt}$)$\times\mathbb{1}_{\text{dist}>100}$			-0.03^{***}			0.02^{***}
			(0.00)			(0.00)
Fixed Effects						
ρ_{ij}	Yes	Yes	Yes	No	No	No
α_{it} and β_{jt}	Yes	Yes	Yes	No	No	No
ρ_{ijo}	No	No	No	Yes	Yes	Yes
α_{iot} and β_{jot}	No	No	No	Yes	Yes	Yes
Pseudo R^2	0.97	0.97	0.97	0.99	0.99	0.99
N	873072	873072	873072	755946	755946	755946

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair-occupation level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. All models incorporate region-occupation-year fixed effects for both residence and workplace and region-pair-occupation fixed effects, enabling an assessment of the travel time reduction and internet access improvement effects within pairs and occupation, holding constant the regions-occupations' time-varying characteristics.

Table 4.7: Effect of travel time and internet access on commuting flows by occupation

However, when controlling for the occupational time varying characteristics by region, α_{iot} and β_{jot}, and for region-pair-occupation time invarying characteristics, ρ_{ijo}, the ranking in the importance of the effect is reversed. This may be primarly due to differences in initial values of commuting flows between occupation. Indeed, we find much more executives commuting over long distances in 1993 than the other type of professions. Still, the increase in commuting flows following travel time reduction is significant.

4.4.3 High-Speed Railways, Internet Access and Complementary Effect

Table 4.8 show the results of the triple difference-in-differences specification. First column displays the effect of travel time reduction evaluated by means of the variable dummy HSR_{ijt}, which equals one after the residence-workplace region-pair has experienced a decrease in travel time at year t, zero otherwise. A reduction in travel time due to an HSR opening is found to increase the amount of commuters about 17.11% on average.

	(1)	(2)	(3)	(4)
$\mathbb{1}_{\text{HSR}_{ijt}}$	0.1711^{***}	0.0166	0.0588^{**}	-0.0124
	(0.0129)	(0.0169)	(0.0251)	(0.0257)
$\mathbb{1}_{\text{HSR}_{ijt}} \times \mathbb{1}_{\text{Internet}_{ijt}}$		0.1590^{***}	0.0808^{***}	
		(0.0119)	(0.0255)	
$\mathbb{1}_{\text{Internet}_{ijt}}$		0.0755^{***}	0.0573^{***}	
		(0.0119)	(0.0189)	
$\mathbb{1}_{\text{Internet}_{ijt}} \times \mathbb{1}_{\text{dist}>100}$			0.1071^{***}	
			(0.0145)	
asinh(internet access$_{ijt}$)				0.0075^{***}
				(0.0027)
asinh(internet access$_{ijt}$)$\times\mathbb{1}_{\text{dist}>100}$				0.0058^{***}
				(0.0007)
HSR$_{ijt}\times$asinh(internet access$_{ijt}$)				0.0056^{***}
				(0.0012)
Fixed Effects				
ρ_{ij}	Yes	Yes	Yes	Yes
α_{it} and β_{jt}	Yes	Yes	Yes	Yes
N	218295	218295	218295	218295
Pseudo R^2	0.99	0.99	0.99	0.99
Average Implied Effect			$\mathbb{1}_{\text{dist}>100}$	
HSR, when Internet $= 0$	17.11%	0.00%	5.88%	-
Internet, when HSR $= 0$	-	7.55%	16.44%	-
HSR, when Internet $= 1$	-	15.90%	13.96%	-
HSR and Internet	-	22.55%	30.40%	-

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. All models incorporate region-year fixed effects for both residence and workplace and region-pair fixed effects, enabling an assessment of the travel time reduction and internet access improvement effects within pairs. The effect of travel time reduction is evaluated by means of the variable dummy HSR_{ijt}, which equals one after the residence-workplace region-pair has experienced a decrease in travel time at year t, zero otherwise. The effect of internet access is evaluated by the means of the variable dummy Internet_{ijt}, which equals one after the residence and workplace regions have internet connection, zero otherwise. I evaluate the implied effects of (1) saved travel time alone, without internet access, which corresponds to this condition: $\Delta\text{HSR}_{ijt} = 1 \wedge \Delta\mathbb{1}(\Delta^{t-1993}\text{internet access}_{ij} > 0) = 0$; (2) internet access alone, without travel time improvement, corresponding to this condition: $\Delta\text{HSR}_{ijt} = 0 \wedge \Delta\mathbb{1}(\Delta^{t-1993}\text{internet access}_{ij} > 0) = 1$; and finally, (3) of the coupled effect of travel time reduction and internet access, following this condition: $\Delta\text{HSR}_{ijt} = 1 \wedge \Delta\mathbb{1}(\Delta^{t-1993}\text{internet access}_{ij} > 0) = 1$.

Table 4.8: HSR and internet access complementarity

The second specification is augmented by introducing a dummy variable, denoted as $\mathbb{1}(\Delta^{t-1993}\text{internet access}_{ij} > 0)$, which signals whether both the residence and workplace regions have internet connection. This dummy takes the value one when both regions are

connected and zero when one of the two is connected. Furthermore, an interaction term is included between the two dummy variables to explore the combined effect of the two technological improvements, which I anticipate both contribute to an increase in commuting flows.

Here, travel time reduction alone is not found to have an effect on the amount of commuters between regions better connected. On the other side, internet connection is found to increase the amount of commuters about 7.55%, whatever the distance between residence and workplace. However, travel time reduction is found to have a positive effect only after residence and workplace regions get connected to the internet, with an average implied effect about 22.55% on commuting flows.

The third specification refines the overall impact of internet access by considering the differential effect at distances below and above 100 kilometers. The results reveal that internet access has an effect of approximately 5.73% at short distances. At longer distances, the impact of internet access increases by 10.71 percentage points, resulting in a total increase of 16.44% in the number of commuters on average. This finding offers initial evidence of the significance of internet access in shaping commuting patterns over long distances, suggesting that individuals opting for a combination of residence and workplace at considerable distances may be engaged in telework activities.

Before internet connection, the effect of travel time reduction is initially estimated to increase commuting flows by 5.88%. Note that it is exactly the estimated increase in commuting flows implied by an average travel time reduction of about 12%, from the results in Table 4.3 (column 4) and Table 4.4 (column 5). This effect is found to be significantly amplified by 24.52 percentage points after internet connection. Consequently, the overall impact of travel time reduction, coupled with internet access, averages to a substantial 30.40% increase in commuting flows between regions that experienced a decrease in travel time due to the opening of a high-speed rail.

The last specification demonstrates that the impact of internet access on commuting flows is both significant and proportional to the percentage increase in the value of bilateral regional internet access. This measurement incorporates internet coverage by ADSL, optic fiber, and their respective loading speeds. The effect of internet access remains higher at distances over 100 kilometers, and the complementary effect of internet access and travel time reduction remains significantly different from zero.

In assessing the impact of internet connection, the dummy chosen was 1 after the variable *Bilateral Internet Access*$_{ijt}$ becomes strictly positive. However, almost every region-pairs get strictly positive values after 2001. Hence, I also test the results using a dummy variable which equals 1 if variable *Bilateral Internet Access*$_{ijt}$ exceeds the median value at time t. To avoid situations where region-pairs fluctuate between values of 1 and 0 due to changes in median values, I retain the dummy variable as 1 if it had a previous value of 1 at time t. This approach expands the pool of observations in the untreated groups among region-pairs. Importantly, this expansion does not alter the results.

4.4.4 HSR, Internet Access and Complementary Effect by Workers' Occupation

Table 4.10 presents the results of the triple difference-in-differences specifications by occupation, while Table 4.9 displays the implied average effects of HSR alone, Internet alone, and HSR and Internet combined on commuting flows.

	(1)	(2)	(3)
			$\mathbb{1}_{\text{dist}>100}$
Executives			
HSR, when Internet $= 0$	10.02%	0.00%	0.00%
HSR, when HSR $= 0$	-	0.00%	0.00%
HSR, when Internet $= 1$	-	9.54%	8.46%
HSR and Internet	-	9.54%	8.46%
Intermediate Professions			
HSR, when Internet $= 0$	0.00%	8.38%	0.00%
Internet, when HSR $= 0$	-	6.81%	-0.78%
HSR, when Internet $= 1$	-	1.95%	0.00%
HSR and Internet	-	8.76%	-0.78%
Employees			
HSR, when Internet $= 0$	23.37%	-5.80%	0.00%
Internet, when HSR $= 0$	-	18.80%	37.27%
HSR, when Internet $= 1$	-	24.38%	14.24%
HSR and Internet	-	43.18%	51.51%
Blue-Collar Workers			
HSR, when Internet $= 0$	25.95%	0.00%	9.53%
Internet, when HSR $= 0$	-	7.72%	27.52%
HSR, when Internet $= 1$	-	23.98%	19.98%
HSR and Internet	-	31.70%	47.50%

The implied average effect of HSR and internet access are computed as the sum of the estimators in Table 4.10 multiplied by 100, applying conditions for travel time reduction and internet access.

Table 4.9: Implied average effects of results in Table 4.10

Results show that improvements in transportation access due to HSR roll-out increase commuting flows by 10% for executives, 0% for workers of intermerdiate profession, 23% for employees and 26% for blue-collar workers. When including internet access dummy and its interaction term with travel time reduction dummy, we find that executives are less impacted than employees and blue-collar workers by HSR and internet improvement. This result is not especially expected, since I identify internet access as a mean to telework and overcoming the commuting costs from large distance between residence and workplace.

In contrast to the model proposed by Gokan et al. (2022), which asserts that unskilled workers do not engage in telework, my findings indicate that certain employees and blue-collar workers do indeed commute over long distances, and more importantly that the impact of reducing travel time is enhanced when coupled with the measure of internet access.

4.4.5 Adjustments margins

In the future version of this paper, I am going to investigate the main adjustment margins, as presented in Section 4.3.3.

	(1)	(2)	(3)
Executives			
$\mathbb{1}_{\text{HSR}_{ijt}}$	0.1002^{***}	0.0097	0.0158
	(0.0245)	(0.0294)	(0.0329)
$\mathbb{1}_{\text{HSR}_{ijt}} \times \mathbb{1}_{\text{Internet}_{ijt}}$		0.0954^{***}	0.0846^{***}
		(0.0178)	(0.0283)
$\mathbb{1}_{\text{Internet}_{ijt}}$		0.0048	0.0019
		(0.0397)	(0.0405)
$\mathbb{1}_{\text{Internet}_{ijt}} \times \mathbb{1}_{\text{dist}>100}$			0.0131
			(0.0208)
Intermediate Professions			
$\mathbb{1}_{\text{HSR}_{ijt}}$	0.0239	0.0838^{***}	0.0478
	(0.0196)	(0.0311)	(0.0311)
$\mathbb{1}_{\text{HSR}_{ijt}} \times \mathbb{1}_{\text{Internet}_{ijt}}$		-0.0643^{**}	0.0081
		(0.0253)	(0.0319)
$\mathbb{1}_{\text{Internet}_{ijt}}$		0.0681^{**}	0.0867^{***}
		(0.0291)	(0.0284)
$\mathbb{1}_{\text{Internet}_{ijt}} \times \mathbb{1}_{\text{dist}>100}$			-0.0945^{***}
			(0.0198)
Employees			
$\mathbb{1}_{\text{HSR}_{ijt}}$	0.2337^{***}	-0.0580^{**}	0.0293
	(0.0207)	(0.0267)	(0.0350)
$\mathbb{1}_{\text{HSR}_{ijt}} \times \mathbb{1}_{\text{Internet}_{ijt}}$		0.3018^{***}	0.1424^{***}
		(0.0178)	(0.0337)
$\mathbb{1}_{\text{Internet}_{ijt}}$		0.1880^{***}	0.1458^{***}
		(0.0423)	(0.0384)
$\mathbb{1}_{\text{Internet}_{ijt}} \times \mathbb{1}_{\text{dist}>100}$			0.2269^{***}
			(0.0215)
Blue-Collar Workers			
$\mathbb{1}_{\text{HSR}_{ijt}}$	0.2595^{***}	0.0220	0.0953^{***}
	(0.0196)	(0.0310)	(0.0252)
$\mathbb{1}_{\text{HSR}_{ijt}} \times \mathbb{1}_{\text{Internet}_{ijt}}$		0.2398^{***}	0.1045^{***}
		(0.0248)	(0.0193)
$\mathbb{1}_{\text{Internet}_{ijt}}$		0.0772^{***}	0.0449^{*}
		(0.0278)	(0.0258)
$\mathbb{1}_{\text{Internet}_{ijt}} \times \mathbb{1}_{\text{dist}>100}$			0.2303^{***}
			(0.0153)
Fixed Effects			
ρ_{ijo}	Yes	Yes	Yes
α_{iot} and β_{jot}	Yes	Yes	Yes
N	755946	755946	755946
Pseudo R^2	0.99	0.99	0.99

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Robust standard errors, clustered at the region-pair-occupation level, are presented in parentheses. The dependent variable is the count of commuters residing in region i and working in region j in year t. All models incorporate region-occupation-year fixed effects for both residence and workplace and region-pair-occupation fixed effects, enabling an assessment of the travel time reduction and internet access improvement effects within pairs and occupation, holding constant the regions-occupations' time-varying characteristics . The effect of travel time reduction is evaluated by means of the variable dummy HSR_{ijt}, which equals one after the residence-workplace region-pair has experienced a decrease in travel time at year t, zero otherwise. The effect of internet access is evaluated by the means of the variable dummy Internet_{ijt}, which equals one after the residence and workplace regions have internet connection, zero otherwise.

Table 4.10: HSR and internet access complementarity by occupation

4.5 Discussion and Conclusion

This research aims to explore the impact of high-speed railways (HSR) on the spatial reallocation of workers, shedding light on the repercussions of improved connectivity on migration and commuting trends. Employing the implementation of HSR as a quasi-natural experiment and utilizing a gravity model with three-way fixed effects, commuting adjustments are modeled in response to changes in travel time between NUTS3-region pairs. The identification of the effect of travel time on commuting flows becomes feasible through its reduction resulting from HSR openings. Additionally, this paper considers the effect of internet access, that can play as an amenity at home and workplace, but could also reduce commuting costs through the lens of telework activities.

Preliminar findings show that workers, including executives, employees, and blue-collar workers, show a growing preference for selecting residences and workplaces in separate and distant locations. This shift is attributed to the improved connectivity of regions facilitated by HSR and internet accessibility. On one hand, results show that after travel time reduction, commuting flows are expected to increase by 6% on average, in the absence of internet access. On the other hand, internet access acts both as a convenient amenity, positively influencing commuting flows across all distances, and as a complement to HSR. The interaction between travel time reduction and internet access indicates that, with internet connectivity, a residence-workplace pair experiencing travel time reduction is expected to see a 14% increase in commuters. Overall, the combined impact of HSR and internet on long-distance commuting (exceeding 100 kilometers) is estimated to be around 30%.

The impact of telework exposure and travel time reduction resulting from the implementation and expansion of France's high-speed rail network may have been magnified following the Covid-19 pandemic in 2020, which triggered significant shifts in commuting behaviors. The nation witnessed an urban exodus, as workers relocated from urban centers to reside further away while still retaining their jobs and embracing telework flexibility, as evidenced by Ramani and Bloom (2021). Subsequently, businesses have widely incorporated telework into contracts and operational processes to optimize time and reduce costs associated with office space and home rentals or price. This adaptation has led firms to downsizing office premises while efficiently managing employee presence on-site or remote work, enabling a rotational arrangement based on the days of the week. Studying the case of France, Bergeaud et al. (2023) indeed show that, as a result of a demand downward shift in office spaces, the depreciation in their value is more pronounced in regions with higher telecommuting exposure. On another hand, households can purchase or rent more spacious homes for the same cost, or homes of equivalent size at a reduced price in smaller urban areas, or even more affordably in rural settings. This rationale has been elaborated upon in a general equilibrium models proposed by Behrens et al. (2021) and Brueckner et al. (2023).

As remote work has become a vital aspect within companies due to the Covid-19 pandemic, SNCF has responded by tailoring its travel offerings to align with this evolving work dynamic and changing commuting patterns. Since September 2021, SNCF has introduced a new subscription option known as the 'annual telework package', called 'max actif package' in 2023. This subscription is designed for individuals who travel 'multiple times a week, following the same route throughout the year'. It particularly caters to those who engage in telecommuting two to three days a week, primarily from Monday to Thursday, allowing up to four daily trips. Practically, this new package grants access to 250 reservations per year at a rate 40% lower

than the traditional "annual package" subscription, which already offers 60% off ticket price with every purchase in addition to a fixed cost. Payment is distributed over twelve monthly installments, and reservations can be made up to two months in advance for both first and second-class tickets.

The French Labor Code has also undergone changes to accommodate the evolving commuting practices. Article L. 3261-2 of the code outlines that employers are required to cover a portion of the subscription ticket expenses incurred by their employees for public transportation between their usual residence and workplace. The court ruled that an employee's home could be defined as the place where they consistently reunite with their family, particularly over weekends, as it reflects the steady and fundamental focus of their personal interests (Cass. soc. 12 Nov. 2020, n° 19-14818). This interpretation applies even if the employee temporarily resides elsewhere during the workweek to reduce commuting distance to their workplace. Employers are obligated to reimburse up to 50% of the ticket cost for the employee, based on the 2nd class fare. To avail this reimbursement, employees are required to provide their transport tickets as evidence to the employer.

Hence, not only has the time cost between distant cities decreased following the HSR expansion, but also the monetary price of commuting due to the previously mentionned arrangements. As of August 2023, examples of monthly SNCF subscription costs for teleworkers include: 347 euros for Lyon-Marseille, 391 euros for Paris-Strasbourg, 348 euros for Paris-Rennes, and 205 euros for Bordeaux-Poitiers. With employers reimbursing 50% of the subscription fee, individuals only bear a cost ranging from 100 to 200 euros per month for these journeys.

Future research will be willing to assess how these new commuting arrangements for teleworkers have impacted commuting patterns. Due to the unavailability of worker data, my analysis is confined to the period up to 2019, which is the most recent data available within the "Panel tous salariés" dataset.[15] It is reasonable to expect that the increase in commuting patterns along high-speed rail routes is likely to be more pronounced than what has been estimated in this study due to decreased travel costs. Nevertheless, the mere existence of these SNCF and work law arrangements underscores the significance of the present study and validates the observed shift in commuting practices attributed to the introduction of high-speed rails. It further underlines the need to tudy those inter-regional commuting patterns as they have the potential to reshape the spatial organization of economic activity and labor markets.

The changing patterns in (inter-regional) commuting can have important implications for regional disparities. Unlike traditional migrants who shift their consumption location along with their move, spending their money in the destination area, inter-regional commuters earn their income at the workplace and expend it in their remote residence. If inter-regional commuters predominantly live in peripheral regions and commute to high-wage urban areas, facilitated by improved connectivity to labor markets through HSR, then the implementation of HSR can contribute to fostering economic development through higher consumption capacity in remote areas, and ultimately reducing regional disparities.

[15]Data for the year 2020 has been accessible since October 23rd, 2023.

Bibliography

Abreu, M. and Öner, Ö. (2020). Disentangling the brexit vote: the role of economic, social and cultural contexts in explaining the uk's eu referendum vote. *Environment and Planning A: Economy and Space*, 52(7):1434–1456.

Acemoglu, D., Gallego, F. A., and Robinson, J. A. (2014). Institutions, human capital, and development. *Annual Review of Economics*, 6(1):875–912.

Adam, H. L., Larch, M., and Stadelmann, D. (2023). Trade agreements and subnational income of border regions. *Economic Inquiry*.

Agrawal, A., Galasso, A., and Oettl, A. (2017). Roads and innovation. *Review of Economics and Statistics*, 99(3):417–434.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Ahmad, S. and Riker, D. (2020). Updated estimates of the trade elasticity of substitution. *US International Trade Commission, Economics Working Paper*.

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., and Sterzi, V. (2018). Dancing with the stars: Innovation through interactions. Technical report, National Bureau of Economic Research.

Alonso, W. (1964). *Location and land use: toward a general theory of land rent*. Harvard university press.

Anderson, J. E., Larch, M., and Yotov, Y. V. (2020). Transitional growth and trade with frictions: A structural estimation framework. *The Economic Journal*, 130(630):1583–1607.

Andersson, D., Berger, T., and Prawitz, E. (2023). Making a market: Infrastructure, integration, and the rise of innovation. *The Review of Economics and Statistics*, 105(2):258–274.

A'Hearn, B. and Venables, A. J. (2013). Regional Disparities: Internal Geography and External Trade. In *The Oxford Handbook of the Italian Economy Since Unification*. Oxford University Press.

Baldwin, R. E. and Martin, P. (2004). Agglomeration and regional growth (chapter 60). In Henderson, J. V. and Thisse, J.-F., editors, *Cities and Geography*, volume 4 of *Handbook of Regional and Urban Economics*, pages 2671–2711. Elsevier.

Barrero, J. M., Bloom, N., and Davis, S. J. (2021). Internet access and its implications for productivity, inequality, and resilience. Technical report, National Bureau of Economic Research.

Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., and Brandt, L. (2020). Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, 115:103124.

Behrens, K., Kichko, S., and Thisse, J.-F. (2021). Working from home: Too much of a good thing?

Bergé, L. et al. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm. Technical report, Department of Economics at the University of Luxembourg.

Bergeaud, A., Eyméoud, J.-B., Garcia, T., and Henricot, D. (2023). Working from home and corporate real estate. *Regional Science and Urban Economics*, 99:103878.

Berger, T. (2019). Railroads and rural industrialization: Evidence from a historical policy experiment. *Explorations in Economic History*, 74:101277.

Bergstrand, J. H., Larch, M., and Yotov, Y. V. (2015). Economic integration agreements, border effects, and distance elasticities in the gravity equation. *European Economic Review*, 78:307–327.

Bergé, L. R. (2015). Network proximity in the geography of research collaboration. Papers in Evolutionary Economic Geography (PEEG) 1507, Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography.

Bernard, A. B., Moxnes, A., and Saito, Y. U. (2020). The geography of knowledge production: Connecting islands and ideas. Technical report, Working paper.

Bircan, C., Javorcik, B., EBRD, O., and Pauly, C. S. (2022). Time after time: Communication costs and inventor collaboration in the multinational firm.

Blonigen, B. A. and Wilson, W. W. (2008). Port efficiency and trade flows*. *Review of International Economics*, 16(1):21–36.

Bonadio, B. (2022). Ports vs. roads: Infrastructure, market access and regional outcomes. *Working Paper*.

Bonadio, B., Dhabi, N. A., Brülhart, M., Cadot, O., and Rais, G. (2023). And there was light: Trade and the development of border regions.

Borjas, G. J. (2014). *Immigration economics*. Harvard University Press.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74.

Boschma, R. and Lambooy, J. (1999). The prospects of an adjustment policy based on collective learning in old industrial regions. *GeoJournal*, 49:391–399.

Boulhol, H. and De Serres, A. (2010). Have developed countries escaped the curse of distance? *Journal of Economic Geography*, 10(1):113–139.

Brakman, S., Garretsen, H., and Marrewijk, C. V. (2009). Economic Geography Within And Between European Nations: The Role Of Market Potential And Density Across Space And Time. *Journal of Regional Science*, 49(4):777–800.

Brakman, S., Garretsen, H., and Schramm, M. (2004). The spatial distribution of wages: Estimating the helpman-hanson model for germany. *Journal of regional science*, 44(3):437–466.

Breinlich, H. (2006). The spatial income structure in the european union—what role for economic geography? *Journal of Economic Geography*, 6(5):593–617.

Brueckner, J. K., Kahn, M. E., and Lin, G. C. (2023). A new spatial hedonic equilibrium in the emerging work-from-home economy? *American Economic Journal: Applied Economics*, 15(2):285–319.

Brülhart, M. (2006). The fading attraction of central regions: an empirical note on core–periphery gradients in western europe. *Spatial Economic Analysis*, 1(2):227–235.

Brülhart, M., Crozet, M., and Koenig, P. (2004). Enlargement and the eu periphery: the impact of changing market potential. *World Economy*, 27(6):853–875.

Brülhart, M., Desmet, K., and Klinke, G.-P. (2020). The shrinking advantage of market potential. *Journal of Development Economics*, 147:102529.

Bruna, F., Lopez-Rodriguez, J., and Faíña, A. (2016). Market potential, spatial dependences and spillovers in european regions. *Regional Studies*, 50(9):1551–1563.

Burda, M. C. and Hunt, J. (2001). From reunification to economic integration: Productivity and the labor market in eastern germany. *Brookings papers on economic activity*, 2001(2):1–92.

Cairncross, F. (1997). The death of distance: How the communications revolution will change our lives. *Cambridge: Harvard Business School Press.*

Camagni, R., Capello, R., Lenzi, C., and Perucca, G. (2023). Urban crisis vs. urban success in the era of 4.0 technologies: Baumol's model revisited. *Papers in Regional Science.*

Catalini, C. (2018). Microgeography and the direction of inventive activity. *Management Science*, 64(9):4348–4364.

Catalini, C., Fons-Rosen, C., and Gaulé, P. (2020). How do travel costs shape collaboration? *Management Science*, 66(8):3340–3360.

Chesbrough, H. (2012). Open innovation: Where we've been and where we're going. *Research-Technology Management*, 55(4):20–27.

Chiquiar, D. (2008). Globalization, regional wage differentials and the Stolper-Samuelson Theorem: Evidence from Mexico. *Journal of International Economics*, 74(1):70–93.

Chiquiar, D. and Hanson, G. H. (2005). International migration, self-selection, and the distribution of wages: Evidence from mexico and the united states. *Journal of political Economy*, 113(2):239–281.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012a). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6):2543–2594.

Combes, P.-P., Duranton, G., Gobillon, L., and Roux, S. (2012b). Sorting and local wage and skill distributions in france. *Regional Science and Urban Economics*, 42(6):913–930.

Combes, P. P., Duranton, G., and Overman, H. G. (2005). Agglomeration and the adjustment of the spatial economy§. *Papers in Regional Science*, 84(3):311–349.

Coughlin, C. C. and Novy, D. (2021). Estimating border effects: The impact of spatial aggregation. *International Economic Review*, 62(4):1453–1487.

Cummings, J. N. and Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social studies of science*, 35(5):703–722.

Daniele, V., Malanima, P., and Ostuni, N. (2018). Geography, market potential and industrialization in italy 1871-2001. *Papers in Regional Science*, 97.

De Fraja, G., Matheson, J., and Rockey, J. (2021). Zoomshock: The geography and local labour market consequences of working from home. *Covid Economics*, (64):1–41.

De Noni, I., Orsi, L., and Belussi, F. (2018). The role of collaborative networks in supporting the innovation performances of lagging-behind european regions. *Research Policy*, 47(1):1–13.

de Palma, A., Picard, N., and Waddell, P. (2007). Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater paris region. *Journal of Urban Economics*, 62(2):204–230.

de Sousa, J. (2012). The currency union effect on trade is decreasing over time. *Economics Letters*, 117(3):917–920.

de Vos, D., van Ham, M., and Meijers, E. J. (2019). Working from home and commuting: Heterogeneity over time, space, and occupations.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Dijkstra, L., Poelman, H., and Rodríguez-Pose, A. (2020). The geography of eu discontent. *Regional Studies*, 54(6):737–753.

Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home? *Journal of public economics*, 189:104235.

Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Donaldson, D. and Hornbeck, R. (2016). Railroads and american economic growth: A "market access" approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Dong, X., Zheng, S., and Kahn, M. E. (2018). The role of transportation speed in facilitating high skilled teamwork. Working Paper 24539, National Bureau of Economic Research.

Ducruet, C. (2020). The geography of maritime networks: A critical review. *Journal of Transport Geography*, 88:102824.

Duranton, G. and Handbury, J. (2023). Covid and cities, thus far. Technical report, National Bureau of Economic Research.

Duranton, G. and Puga, D. (2020). The economics of urban density. *Journal of economic perspectives*, 34(3):3–26.

Duranton, G. and Turner, M. A. (2012). Urban growth and transportation. *Review of Economic Studies*, 79(4):1407–1440.

Eder, J. (2019). Innovation in the periphery: A critical survey and research agenda. *International Regional Science Review*, 42(2):119–146.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from china's national trunk highway system. *Review of Economic Studies*, 81(3):1046–1070.

Faist, T. (2000). The volume and dynamics of international migration and transnational social spaces.

Fallah, B., Partridge, M., and Olfert, M. (2009). New economic geography and us metropolitan wage inequality. *Journal of Economic Geography*, 11.

Fally, T., Paillacar, R., and Terra, C. (2010). Economic geography and wages in Brazil: Evidence from micro-data. *Journal of Development Economics*, 91(1):155–168.

Feng, Q., Chen, Z., Cheng, C., and Chang, H. (2023). Impact of high-speed rail on high-skilled labor mobility in china. *Transport Policy*, 133:64–74.

Fingleton, B. (2008). Competing models of global dynamics: evidence from panel models with spatially correlated error components. *Economic Modelling*, 25(3):542–558.

Fontagné, L., Guimbard, H., and Orefice, G. (2022). Tariff-based product-level trade elasticities. *Journal of International Economics*, 137:103593.

Frankel, J. A. and Romer, D. H. (1999). Does trade cause growth? *American Economic Review*, 89(3):379–399.

Frenken, K., Hoekman, J., Kok, S., Ponds, R., Oort, F., and Vliet, J. (2009). *Death of Distance in Science? A Gravity Approach to Research Collaboration*, pages 43–57.

Frensch, R., Fidrmuc, J., and Rindler, M. (2023). Topography, borders, and trade across europe. *Journal of Comparative Economics*.

Fujita, M., Krugman, P., and Venables, A. J. (1999). *The spatial economy: Cities, regions, and international trade*. MIT press.

Gallego, J., Rubalcaba, L., and Suárez, C. (2013). Knowledge for innovation in europe: The role of external knowledge on firms' cooperation strategies. *Journal of Business Research*, 66(10):2034–2041.

Gao, Y. and Zheng, J. (2020). The impact of high-speed rail on innovation: An empirical test of the companion innovation hypothesis of transportation improvement with china's manufacturing firms. *World Development*, 127:104838.

Gaspar, J. and Glaeser, E. (1998). Information technology and the future of cities. *Journal of Urban Economics*, 43(1):136–156.

Gaulier, G. and Zignago, S. (2010). Baci: International trade database at the product-level. the 1996-2020 version. Working Papers 2010-23, CEPII.

Gennaioli, N., LaPorta, R., de Silanes, F. L., and Shleifer, A. (2013). Human capital and regional development. *Quarterly Journal of Economics*, 128(1):105–164.

Gennaioli, N., Porta, R. L., Silanes, F. L. D., and Shleifer, A. (2014). Growth in regions. *Journal of Economic Growth*, 19(3):259–309.

Giroud, X., Lenzu, S., Maingi, Q., and Mueller, H. (2021). Propagation and amplification of local productivity spillovers. Technical report, National Bureau of Economic Research.

Gokan, T., Kichko, S., Matheson, J., and Thisse, J.-F. (2022). How the rise of teleworking will reshape labor markets and cities.

Guellec, D. and van Pottelsberghe de la Potterie, B. (2001). The internationalisation of technology analysed with patent data. *Research Policy*, 30(8):1253–1266.

Guerriero, M. (2019). *The labor share of income around the world: Evidence from a panel dataset*. Springer.

Guirao, B., Campa, J. L., and Casado-Sanz, N. (2018). Labour mobility between cities and metropolitan integration: The role of high speed rail commuting in spain. *Cities*, 78:140–154.

Haas, A. and Osland, L. (2014). Commuting, migration, housing and labour markets: Complex interactions.

Hallépée, S. and Mauroux, A. (2019). Quels sont les salariés concernés par le télétravail? *Dares Analyses*, 051:11.

Hanley, D., Li, J., and Wu, M. (2022). High-speed railways and collaborative innovation. *Regional Science and Urban Economics*, 93:103717.

Hanson, G. H. (2005). Market potential, increasing returns and geographic concentration. *Journal of International Economics*, 67(1):1–24.

Harris, C. D. (1954). The market as a factor in the localization of industry in the united states. *Annals of the Association of American Geographers*, 44(4):315–348.

Head, K. and Mayer, T. (2004). Market potential and the location of japanese investment in the european union. *Review of Economics and Statistics*, 86(4):959–972.

Head, K. and Mayer, T. (2006). Regional wage and employment responses to market potential in the EU. *Regional Science and Urban Economics*, 36(5):573–594.

Head, K. and Mayer, T. (2011). Gravity, market potential and economic development. *Journal of Economic Geography*, 11(2):281–294.

Head, K. and Mayer, T. (2014). Gravity Equations: Workhorse,Toolkit, and Cookbook. In Gopinath, G., Helpman, ., and Rogoff, K., editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, chapter 0, pages 131–195. Elsevier.

Head, K. and Ries, J. (2001). Increasing returns versus national product differentiation as an explanation for the pattern of u.s.-canada trade. *American Economic Review*, 91(4):858–876.

Hering, L. and Paillacar, R. (2016). Does access to foreign markets shape internal migration? evidence from brazil. *World Bank Economic Review*, 30(1):78–103.

Hering, L. and Poncet, S. (2010). Market access and individual wages: Evidence from china. *The Review of Economics and Statistics*, 92(1):145–159.

Heuermann, D. F. and Schmieder, J. F. (2019). The effect of infrastructure on worker mobility: evidence from high-speed rail expansion in germany. *Journal of economic geography*, 19(2):335–372.

Hoekman, J., Frenken, K., and Oort, F. (2009). The geography of collaborative knowledge production in Europe. *The Annals of Regional Science*, 43(3):721–738.

Hornbeck, R. and Rotemberg, M. (2021). Railroads, market access, and aggregate productivity growth. *University of Chicago Booth School of Business, mimeo*.

Hotelling, H. (1929). Stability in competition. *The Economic Journal*, 39(153):41–57.

Hummels, D. (2007). Transportation costs and international trade in the second era of globalization. *Journal of Economic perspectives*, 21(3):131–154.

Jacks, D. S. and Novy, D. (2018). Market Potential and Global Growth over the Long Twentieth Century. *Journal of International Economics*, 114(C):221–237.

Jacobs, J. (1969). The economy of cities. *Random House, New York*.

Jaffe, A. B. (1986). Technological opportunity and spillovers of r & d: Evidence from firms' patents, profits, and market value. *The American Economic Review*, 76(5):984–1001.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Jarvenpaa, S. L. and Leidner, D. E. (1998). Communication and trust in global virtual teams. *Journal of computer-mediated communication*, 3(4):JCMC346.

Kang, M., Li, Y., Zhao, Z., Song, M., and Yi, J. (2023). Travel costs and inter-city collaborative innovation: Evidence of high-speed railway in china. *Structural Change and Economic Dynamics*, 65:286–302.

Karahasan, B. C., Dogruel, F., and Dogruel, A. S. (2016). Can market potential explain regional disparities in developing countries? evidence from turkey. *The Developing Economies*, 54(2):162–197.

Kennan, J. and Walker, J. R. (2011). The effect of expected income on individual migration decisions. *Econometrica*, 79(1):211–251.

Kerr, W. R. and Robert-Nicoud, F. (2020). Tech clusters. *Journal of Economic Perspectives*, 34(3):50–76.

Koh, Y., Li, J., and Xu, J. (2022). Subway, collaborative matching, and innovation. *Review of Economics and Statistics*, pages 1–45.

Koser, K. (2007). *International migration: A very short introduction*. Oxford University Press.

Kosfeld, R. and Eckey, H.-F. (2010). Market access, regional price level and wage disparities: the german case. *Jahrbuch für Regionalwissenschaft*, 30(2):105–128.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, 70(5):950–59.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, 99(3):483–99.

Lai, H. and Trefler, D. (2002). The gains from trade with monopolistic competition: specification, estimation, and mis-specification.

Lanjouw, J. O. and Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The economic journal*, 114(495):441–465.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, pages 319–333.

Li, C., Zhou, Q., and Chen, S. (2022). Bringing minds together: High-speed railways, team building, and innovation collaboration. *China & World Economy*, 30(6):34–58.

Limão, N. and Venables, A. J. (2001). Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review*, 15(3):451–479.

Liu, S. and Su, Y. (2023). The effect of working from home on the agglomeration economies of cities: Evidence from advertised wages. *Available at SSRN 4109630*.

López-Rodríguez, J. and Andrés Faíña, J. (2006). Does distance matter for determining regional income in the european union? an approach through the market potential concept. *Applied Economics Letters*, 13(6):385–390.

López-Rodríguez, J., Márquez, M. A., and Faiña, A. (2008). Economic geography and spatial wage structure in spain. *REAL Discussion Papers 08-T*, 4.

Malgouyres, C., Mayer, T., and Mazet-Sonilhac, C. (2021). Technology-induced trade shocks? evidence from broadband expansion in france. *Journal of International Economics*, 133:103520.

Marchiori, M. and Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3-4):539–546.

Marshall, A. (1890). *Principles of economics, by Alfred Marshall.* Macmillan and Company.

Mayer, T. and Trevien, C. (2017). The impact of urban public transportation evidence from the paris region. *Journal of Urban Economics*, 102:1–21.

McCallum, J. (1995). National borders matter: Canada-us regional trade patterns. *The American Economic Review*, 85(3):615–623.

Melitz, M. J. and Ottaviano, G. I. (2008). Market size, trade, and productivity. *The review of economic studies*, 75(1):295–316.

Miguelez, E. (2019). Collaborative patents and the mobility of knowledge workers. *Technovation*, 86:62–74.

Mion, G. (2004). Spatial externalities and empirical analysis: the case of Italy. *Journal of Urban Economics*, 56(1):97–118.

Mion, G. and Naticchioni, P. (2009). The spatial sorting and matching of skills and firms. *Canadian Journal of Economics/Revue canadienne d'économique*, 42(1):28–55.

Monte, F., Porcher, C., and Rossi-Hansberg, E. (2023). Remote work and city structure. *American Economic Review*, 113(4):939–981.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–3890.

Montobbio, F. and Sterzi, V. (2013). The globalization of technology in emerging markets: A gravity model on the determinants of international patent collaborations. *World Development*, 44:281–299.

Morescalchi, A., Pammolli, F., Penner, O., Petersen, A. M., and Riccaboni, M. (2015). The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy*, 44(3):651–668.

Moretti, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–3375.

Morrison, G., Riccaboni, M., and Pammolli, F. (2017). Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific data*, 4(1):1–21.

Niebuhr, A. (2006). Market access and regional disparities: New economic geography in europe. *The Annals of Regional Science*, 40(2):313–334.

Niebuhr, A., Granato, N., Haas, A., and Hamann, S. (2012). Does labour mobility reduce disparities between regional labour markets in germany? *Regional Studies*, 46(7):841–858.

Nilles, J. M. (1991). Telecommuting and urban sprawl: mitigator or inciter? *Transportation*, 18(4):411–432.

Ottaviano, G. I. and Pinelli, D. (2006). Market potential and productivity: Evidence from finnish regions. *Regional Science and Urban Economics*, 36(5):636–657.

Paredes, D. (2013). The role of human capital, market potential and natural amenities in understanding spatial wage disparities in chile. *Spatial Economic Analysis*, 8(2):154–175.

Pauly, S. and Stipanicic, F. (2022). The creation and diffusion of knowledge: Evidence from the jet age.

Perlman, E. R. et al. (2016). Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century america. *Work. Pap., Boston Univ.*

Picci, L. (2010). The internationalization of inventive activity: A gravity model using patent data. *Research Policy*, 39(8):1070–1081.

Pires, A. J. G. (2006). Estimating krugman's economic geography model for the spanish regions. *Spanish Economic Review*, 8(2):83–112.

Porter, M. E. (1990). The competitive advantage of nations. *New York (NY): Free Press.*

Portugal-Perez, A. and Wilson, J. S. (2012). Export Performance and Trade Facilitation Reform: Hard and Soft Infrastructure. *World Development*, 40(7):1295–1307.

Ramani, A. and Bloom, N. (2021). The donut effect of covid-19 on cities. Technical report, National Bureau of Economic Research.

Redding and Venables, A. J. (2004). Economic geography and international inequality. *Journal of International Economics*, 62(1):53–82.

Redding, S. J. (2022). Chapter 3 - trade and geography. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics: International Trade, Volume 5*, volume 5 of *Handbook of International Economics*, pages 147–217. Elsevier.

Redding, S. J. and Sturm, D. M. (2008). The costs of remoteness: Evidence from german division and reunification. *American Economic Review*, 98(5):1766–1797.

Reshef, A. and Santoni, G. (2023). Are your labor shares set in beijing? the view through the lens of global value chains. *European Economic Review*, 155:104459.

Rodríguez-Pose, A., Lee, N., and Lipp, C. (2020). Golfing with trump: Social capital, decline, inequality, and the rise of populism in the us.

Romalis, J. (2007). Nafta's and cusfta's impact on international trade. *The Review of Economics and Statistics*, 89(3):416–435.

Schulz, R., Watson, V., and Wersing, M. (2023). Teleworking and housing demand. *Regional Science and Urban Economics*, page 103915.

Shearmur, R. (2012). Are cities the font of innovation? a critical review of the literature on cities and innovation. *Cities*, 29:S9–S18.

Silva, J. M. C. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Tinbergen, J. (1962). Shaping the world economy; suggestions for an international economic policy.

Tsiachtsiras, G. (2022). Changing the perception of time: Railroads, access to knowledge and innovation in nineteenth century france. *Available at SSRN 4297205*.

Tsiachtsiras, G., Yin, D., Miguelez, E., and Moreno, R. (2022). Trains of thought: High-speed rail and innovation in china. *Available at SSRN 4280769*.

Tóth, G., Juhász, S., Elekes, Z., and Lengyel, B. (2021). Repeated collaboration of inventors across european regions. *European Planning Studies*, 0(0):1–21.

Van Houtum, H. and Van Der Velde, M. (2004). The power of cross-border labour market immobility. *Tijdschrift voor economische en sociale geografie*, 95(1):100–107.

Visser, E.-J. and Boschma, R. (2004). Learning in districts: Novelty and lock-in in a regional context. *European Planning Studies*, 12(6):793–808.

von Proff, S. and Brenner, T. (2014). The dynamics of inter-regional collaboration: an analysis of co-patenting. *The Annals of Regional Science*, 52:41–64.

Wang, F., Wei, X., Liu, J., He, L., and Gao, M. (2019). Impact of high-speed rail on population mobility and urbanisation: A case study on yangtze river delta urban agglomeration, china. *Transportation research part A: policy and Practice*, 127:99–114.

Weber, A. and Friedrich, C. J. (1929). Theory of the location of industries. *(No Title)*.

Wessel, P. and Smith, W. (1996). A global, self-consistent, hierarchical, high-resolution shoreline database. *Journal of Geophysical Research*, 101:8741–8743.

Yao, L. and Li, J. (2022). Intercity innovation collaboration and the role of high-speed rail connections: evidence from chinese co-patent data. *Regional Studies*, 56(11):1845–1857.

Yotov, Y. V., Piermartini, R., Larch, M., et al. (2016). *An advanced guide to trade policy analysis: The structural gravity model*. WTO iLibrary.

Zou, W., Chen, L., and Xiong, J. (2021). High-speed railway, market access and economic growth. *International Review of Economics & Finance*, 76:1282–1304.

Özgüzel, C. (2022). Agglomeration effects in a developing economy: evidence from Turkey. *Journal of Economic Geography*. lbac035.