Published online ahead of print April 4, 2013

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–23
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

informs

http://dx.doi.org/10.1287/mnsc.1120.1700

Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity

Jasjit Singh
INSEAD, Singapore 138676, jasjit.singh@insead.edu

Matt Marx
MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, mmarx@mit.edu

Geographic localization of knowledge spillovers is a central tenet in multiple streams of research. However, prior work has typically examined this phenomenon considering only one geographic unit—country, state, or metropolitan area—at a time and has rarely accounted for spatial distance. We disentangle these multiple effects by using a regression framework employing choice-based sampling to estimate the likelihood of citation between random patents. We find both country and state borders to have independent effects on knowledge diffusion beyond what just geographic proximity in the form of metropolitan collocation or shorter within-region distances can explain. An identification methodology comparing inventor-added and examiner-added citation patterns points to an even stronger role of political borders. The puzzling state border effect remains robust on average across analyses, though it is found to have waned with time. The country effect has, in contrast, not only remained robust but even strengthened over time.

Key words: knowledge spillovers; borders; distance; economic geography; patent citation; innovation; institutions
History: Received March 25, 2011; accepted November 24, 2012, by Kamalini Ramdas, entrepreneurship and innovation. Published online in *Articles in Advance*.

1. Introduction

Empirically establishing the microfoundations of industrial agglomeration is a key focus in multiple streams of research. Ever since the seminal work of Marshall (1920), scholars have studied not just exogenous locational factors but also three endogenous mechanisms for why agglomeration takes place: benefits from labor pooling, efficiency gains from collocation of industries with input–output relationships, and localized knowledge spillovers.[1] Of these, knowledge spillovers have generated the most scholarly attention in recent years, perhaps because they are seen as critical for innovation and new value creation in an increasingly knowledge-intensive economy. In this study, we take a closer look at various geographic elements in shaping such spillovers, distinguishing between the roles played by political borders (at the national or state level) versus simply spatial proximity.

Although several studies have documented localization of knowledge spillovers, the geographic levels most relevant for this phenomenon remain unclear, given the approaches employed. For instance, a significant body of empirical work has studied only country-level spillovers (Branstetter 2001; Keller 2002; Jaffe and Trajtenberg 2002, Chap. 7; Singh 2007), with the findings from such studies then being justification for assumptions used in theoretical models of economic growth of nations (Romer 1990, Grossman and Helpman 1991). Others have taken political borders at a less aggregate level—states—as the geographic unit of interest (Jaffe 1989, Audretsch and Feldman 1996, Almeida and Kogut 1999, Rosenthal and Strange 2001), but mostly without consideration of national borders or geographic distance. The typical focus on just one of the borders arises for two reasons—one practical and the other more theoretical. The practical reason has been that it is hard to obtain precise measures for geographic proximity, so national or state borders can be seen as a convenient way of measuring it indirectly. The more substantive reason has been a belief that political borders are important over and above any spatial proximity effects, for example, caused by institutional differences. In either case, however, the extent to which political borders and geographic proximity measure the same thing or not has not been typically tested.

[1] See Ellison et al. (2010) for a recent study that rigorously demonstrates Marshallian mechanisms.

A glaring gap in the literature therefore remains—few studies have attempted to rigorously disentangle the effects operating at different geographic levels, providing limited guidance regarding the exact geographic scope of knowledge spillovers. For example, the fact that within-country knowledge spillovers are found to be more intense than those across countries might simply reflect an aggregation of state-level or metropolitan-level mechanisms. Similarly, interpretation of state-level localization findings is also unclear, as these too might be driven by effects operating more locally and are open to criticisms to the effect that "state boundaries are a very poor proxy for the geographical units within which knowledge ought to circulate" (Breschi and Lissoni 2001, p. 982). Perhaps motivated by such ambiguities, or by criticism like Krugman's (1991, p. 43) remark that "states aren't really the right geographic units" in economic analysis, recent research appears to have renewed focus on exploring agglomeration at less coarsely defined geographic levels.

The economic geography literature has a long tradition of emphasizing a link between localized knowledge spillovers and urban growth (Jacobs 1969, Glaeser et al. 1992, Saxenian 1994). More recently, studies such as Rosenthal and Strange (2003), Singh (2005), and Arzaghi and Henderson (2008) have carefully established agglomeration effects related to such spillovers as indeed being particularly strong over short distances within a city (a few miles or even less). This research is complemented by the broader literature emphasizing the role knowledge spillovers through spin-offs—also often geographically localized—can play in industry and regional growth (Gompers et al. 2005, Klepper and Sleeper 2005, Agarwal et al. 2007). Advances are now being made toward formal models that can capture microfoundations of the geographic scope of geographic clusters like Silicon Valley and calibrating these models against real data (Kerr and Kominers 2010).

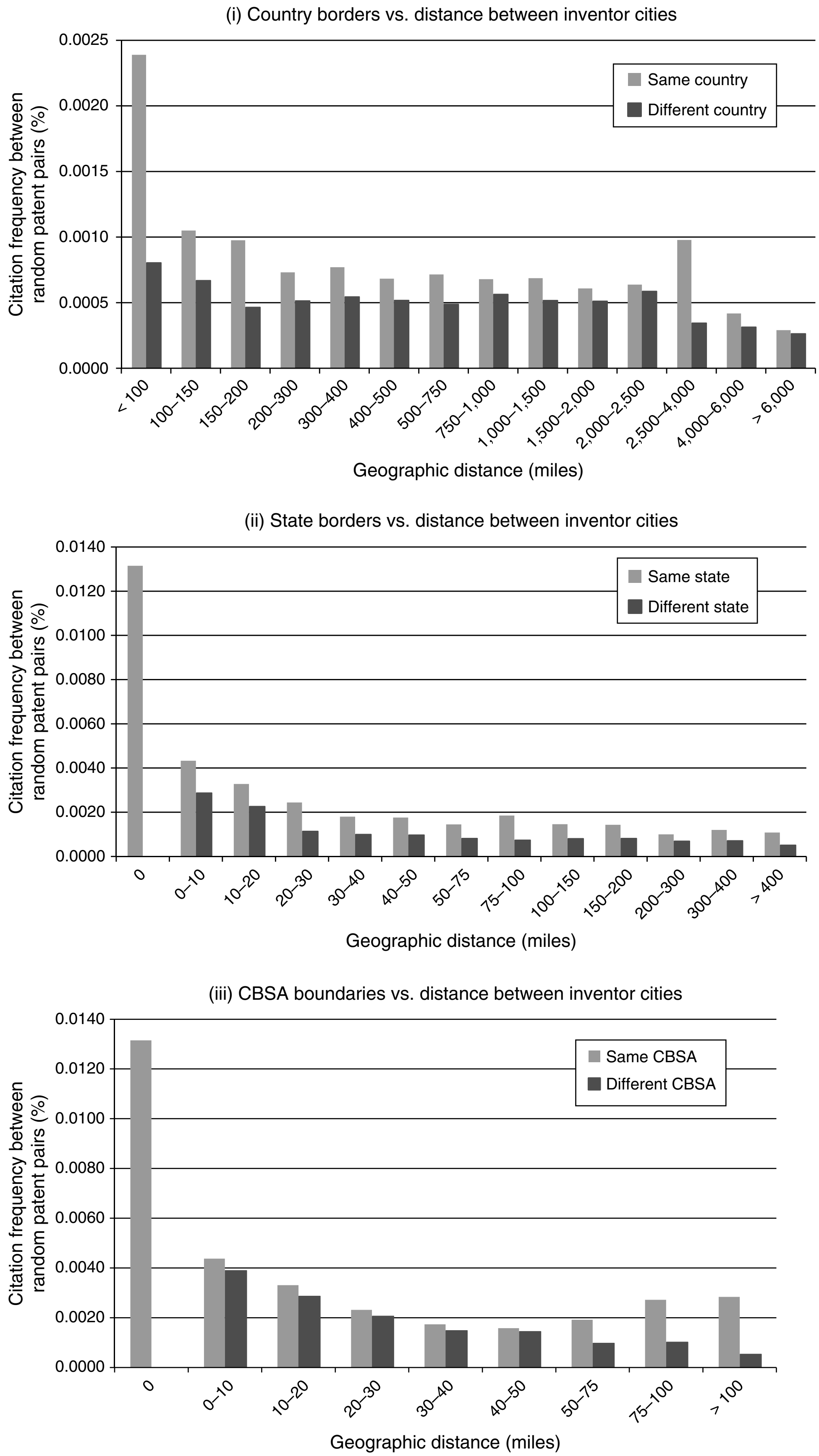
Despite these rich yet distinct bodies of work examining different geographic levels, few studies related to knowledge spillovers have considered different levels of border and proximity effects *simultaneously* in an attempt to unpack the true contribution of each. In addition to not separating the country, state, and metropolitan effects from one another, most studies do not separately identify these from the role of spatial distance either. It is therefore unclear whether to interpret prior findings simply as reflecting that "distance matters" or as borders also having an important and independent role. Even studies that do consider multiple geographic levels, such as the path-breaking paper by Jaffe et al. (1993), analyze these different geopolitical units *separately* and again do not account for precise spatial distance. When at least some distance-based measures have been employed, most have still been too aggregate to disentangle the precise geographic effects of interest. For example, although Keller (2002) employs data on distance between capital cities of countries, he does not consider different within-country distances. Likewise, Peri (2005) extends this to consider distances between different pairs of states, but still does not distinguish city-to-city distances within a state. A forthcoming paper by Belenzon and Schankerman (2013) is an exception that does simultaneously consider geographic distance and collocation even within a state. However, their analysis is based only on knowledge originating from universities, which composes a tiny fraction of the overall knowledge created in the economy and plays a relatively minor role in overall knowledge diffusion patterns.

We analyze the role of borders versus proximity in diffusion of a large knowledge base represented by a set of 631,586 patents, employing recent empirical advancements in studying knowledge diffusion and documenting long-term trends in both within-country and within-state localization effects after multiple geographic effects are all accounted for together. Given the role that specific assumptions regarding the geographic scope of knowledge spillovers can play in research areas as diverse as regional and international economics, business strategy, and technological innovation and entrepreneurship, we see our study as fulfilling a need to dig deeper into the geography of knowledge spillovers in a manner analogous to advances in the literature on cross-regional trade—a field that has made much more progress in examining the borders versus proximity question at both the country level (e.g., McCallum 1995, Anderson and Wincoop 2003) and the state level (e.g., Wolf 2000; Hillberry and Hummels 2003, 2008).

Our empirical approach builds on the increasingly sophisticated use of patent citations to measure diffusion of knowledge. As a motivation for examining the above questions using patent citation data, Figure 1 provides simple graphical evidence regarding the geographic pattern of interfirm citations to patents filed during 1975–2004 by inventors based in the United States.[2] Three observations are worth making. First, the likelihood of citation between random pairs of patents decreases with geographic distance. Second, the citation likelihood is greater within country borders than across, within state borders than across, and within metropolitan area boundaries (measured

[2] These charts were constructed using our data set and employed in our regression analysis. Although our data set was derived using stratified sampling from the population of patent pairs, we calculated the summary statistics by appropriately weighting each observation so that Figure 1 represents true population characteristics.

Figure 1 **Graphical Depiction of the Role of Geography in Patent Citation Likelihood**



using "CBSA" definitions explained later) than across. Third, the national and state border effects seem to be only partly explained by geographic proximity: there seems to be an independent "border effect" within each of the buckets for spatial distance (a finding that continues to hold even with further refining the distance buckets used in the figure).

Figure 1 is, however, just a depiction of summary statistics and does not account for a number of empirical issues addressed in more formal analyses (reported later as a part of this study). In our preferred approach, we employ regression models to run a "horse race" among different geographic variables to isolate the level at which localization of knowledge spillovers operates most prominently. Specifically, we construct a data set of patent pairs (representing actual or potential patent citations) using choice-based sampling and estimate a "citation function" modeling the likelihood of citations between random patents. Our framework departs from previous studies by making no ex ante assumptions about the most appropriate geographic unit of analysis. Instead, it allows us to simultaneously account for collocation of the source and destination of knowledge within the same country, state, or metropolitan area as well as for fine-grained spatial distance.

Consistent with prior work, separate analyses we conduct at the national, state, and metropolitan levels exhibit spillover localization at all levels. Importantly, simultaneously accounting for all of these shows how individual analyses overstate their respective importance. We also extend the analysis to first parametrically control for distance and then employ a set of nonparametric indicator variables to more flexibly capture the more nuanced effects of distance. A key finding is that robust border effects are seen even after accounting for geographic proximity. In other words, we continue to see independent country border and state border effects even after carefully controlling for the effect of metropolitan area collocation and the decaying of spillover intensity with spatial distance.

The same-country localization of knowledge spillovers turns out to be several times stronger than the same-state effect. One could view this large and robust national border effect as perhaps in line with expectation, given the well-documented linguistic, cultural, institutional, and economic differences among countries (Coe et al. 2009). However, time-trend analysis also reveals a more surprising finding: a strengthening of the same-country effect over time despite the accepted trend toward globalization and technological advances that supposedly smooth cross-border communication.

We find the state border effect even more puzzling. The finding turns out not to be driven by just one or two specific states (like California) or sectors (like computers or communication technologies). The result is seen even in subsamples comprised of cited patents close to state borders, so it is also not driven just by those in the interior of states. In fact, significant state border effects are found even in a conservative test where metropolitan effects are completely isolated by considering only patents and (potential) citations arising within a subset of metropolitan areas that span state borders. We also analyze trends over time and find that—in contrast to country borders—the state border effect weakens considerably over the 30-year time period from which our cited patent sample is drawn.

To further boost confidence in our findings, we confront two key challenges inherent in using patent citations to measure knowledge spillover localization. First, citation patterns are determined in part by technological relationships, which cannot be perfectly captured by any formal classification system (Thompson and Fox-Kean 2005). Second, some citations are added by patent examiners, not inventors, and the extent to which the two represent spillovers likely differs (Alcacer and Gittelman 2006). To address these concerns, we first examine the robustness of our prior findings to using only inventor-added citations. We then employ an identification strategy (motivated by Thompson 2006) that calculates true geographic effects as the difference between localization estimates for subsamples of citations added by inventors versus patent examiners. Border effect findings remain robust, even though this approach weakens the effect of proximity, further highlighting the importance of borders beyond pure geographic proximity in shaping knowledge diffusion patterns.

2. Empirical Approach

2.1. Constructing a Patent-Based Data Set

Although citation-based measures are noisy in capturing true knowledge flows, surveys of inventors have established that citations—especially in large samples—do capture knowledge flows meaningfully (Jaffe and Trajtenberg 2002, Duguet and MacGarvie 2005). Even assuming that citations do correctly capture knowledge flows, it is not possible to decipher when a given citation represents a "spillover," that is, a true externality for which the receiver does not fully pay. Nevertheless, we follow the prevalent view that citations do at least partly represent spillovers and for the rest often represent benefits the receiver gets in the form of "gains from trade," even in cases where they represent purely market transactions.

Our data set is based on U.S. Patent and Trademark Office (USPTO) patents with application years 1975–2009. To have at least a five-year citation window, we restrict the set to cited patents until 2004, with citing patents going all the way until 2009. Because recent literature (Alcacer and Gittelman 2006, Thompson 2006) has emphasized the distinction between citations added by the inventors themselves and those added by patent examiners working for USPTO, we also keep track of this information when available (2001 onward) and use it to complement our full-sample analyses. Patent data also include inventors' city, state, and country of residence. Consistent state identification is available only for patents originating in the United States, so we restrict the cited patent sample (but not the citing patent sample) to those filed by U.S. inventors. Our calculation of geographic distances relies on data from Lai et al. (2009) that map cities where inventors live to latitudes and longitudes.[3]

We also map these cities to core based statistical areas (CBSAs), a U.S. Office of Management and Budget definition effective as of 2003 and using data from the 2000 Census. A CBSA refers to a county (or local equivalent) with a population center of at least 10,000 plus surrounding counties within reasonable commuting distance. CBSAs with a population center of more than 50,000 are designated as metropolitan statistical areas (MSAs), or micropolitan otherwise.[4] The CBSA definitions enable us to classify considerably more patents than with earlier definitions based on the 1990 Census (96.3% versus 84.7%).[5] Our results are, however, robust to either classification system.

Before proceeding to construct a sample of patent pairs representing actual or potential citations, we restrict the cited patent sample to only patents whose geographic origin is unambiguously defined in order to avoid making arbitrary assumptions in trying to resolve locational ambiguity of a knowledge source. In other words, we exclude patents from geographically dispersed inventor teams, even though these might be an interesting (but different) topic to study. We also omit patents not assigned to any organization, as well as those assigned to nonfirm sources (such as universities and government bodies), as the focus of this study is to examine interfirm diffusion of knowledge. In the end, all the steps mentioned above yield a set of 631,586 potentially cited patents as sources of knowledge.

2.2. Constructing a Matched Sample of Actual and Potential Citations

For each cited patent mentioned above, we collect data on all citations received during a 10-year window since its application and drop all within-firm citations. As a highly influential study by Jaffe et al. (1993) (hereafter, JTH) points out, just calculating collocation frequency within pairs of patents involved in realized citations would not suffice for establishing geographic localization of knowledge. Instead, what is needed is an appropriate control sample of potential (but unrealized) citations to establish a benchmark level of expected collocation, given the existing geographic distribution of technological activity. To facilitate a comparison of our subsequent analysis with the JTH method, we therefore also start with their approach of matching each citing patent to a random control patent with the same three-digit technology class and application year as the original citing patent (but not from the same organization as the focal cited patent and also not actually citing it). Like Jaffe et al. (1993), we drop the small fraction of citations for which no match is found. This leads to a balanced sample of 4,007,217 realized citations (based on 631,586 cited patents) and exactly as many unrealized matched control citations.

The above JTH-style sample allows us to compare the extent of geographic collocation of the source and destination for the original sample of citations and the sample of corresponding control pairs, in turn using the country, state, and metropolitan area as the geographic units of analysis in three *different* sets of calculations. Although it is a useful starting point, this approach is not well suited to directly addressing our key question: How much do national or state borders per se constrain knowledge flow, as opposed to the corresponding effects being manifestations of mechanisms operating at more local levels (such as city or CBSA) or driven purely by spatial distance? Our preferred approach for answering these questions is a regression framework that can simultaneously examine the effect of different geographic levels.

2.3. A Regression Framework for Estimating Citation Likelihood

With collocation within a certain predefined geographic unit as the dependent variable in the JTH model, one cannot easily examine multiple geographic levels at the same time. One could try to ascertain the relative importance of different geographic levels by somehow comparing the findings

[3] Our distance data are restrictive in two ways. First, because we have only a single latitude and longitude coordinate per city, we cannot calculate distances between inventors within a city or even be completely precise about distances between those in different cities. Second, USPTO data contain the city of residence of the inventor, which might not coincide with the city where the work was done. But this is still the best available proxy for an invention's geographical origin, as the assignee address typically refers only to the overall firm's registered office or headquarters.

[4] A cruder approach (that has been employed in the past) could be to designate one single "phantom MSA" for each state to handle cases where an inventor does not fall within an actual MSA. However, doing this would confound metropolitan effects with state effects and is therefore inappropriate for the research question of interest here.

[5] Additional details on CBSA definitions are at https://www.census.gov/geo/www/2010census/gtc/gtc_cbsa.html (accessed August 2012).

across models; however, this would likely remain a statistically complex and unsatisfactory exercise. We instead rely on a regression framework that estimates the likelihood of citation between two random patents, making the existence of a citation between a pair of patents the dependent variable and employing the entire set of geography-related variables *simultaneously* as explanatory variables in a single model.[6]

Our citation-level regression framework has the added advantage of flexibility in modeling technological relatedness between patents, allowing multiple levels of technological granularity to be considered at once. This addresses a criticism previous studies have faced in choosing a specific technological granularity when constructing a JTH-style control sample. As Thompson and Fox-Kean (2005) and Henderson et al. (2005) discuss, one faces a dilemma in using matching: the three-digit technology match commonly employed might be too crude to capture all relevant technological relationships, but using a finer classification could result in selection bias because a match would not be found for most of the sample. Both papers suggest a regression approach that simultaneously accounts for technological relatedness at multiple levels of granularity.[7]

A seemingly straightforward extension of the JTH methodology might be to employ a regression approach using a JTH-style matched sample in a (logit or probit) regression model, wherein the existence of a citation between a pair of patents is taken as the dichotomous dependent variable. However, this would imply that the matching procedure was in effect used to carry out sampling based on the dependent variable in the first place, because the JTH method draws a "zero" (unrealized citation) corresponding to each "one" (actual citation). This needs to be taken into account to get estimates truly representative of the population. Further, the potentially citing patents used in constructing the control pairs are drawn only from technology classes and years from which citations to the cited patent actually exist, ignoring the population of potentially citing patents from the remaining technology classes and years. As the appendix explains, this can further bias the results. Here, building on Singh (2005), we describe a microlevel citation regression framework that ameliorates this issue.

[6] Our methodology builds on studies such as Sorenson and Fleming (2004) and Singh (2005) that also model the citation likelihood between patents in a regression framework, though to study different research questions.

[7] This does not fully address the issue that no technological classification system—however finely defined—can perfectly capture true technological relationships between patents. We address this concern later in the paper by extending our JTH-style as well as a regression analysis using an approach motivated by Thompson (2006).

Before discussing how to extend our matched sample for citation-level regression analysis, for exposition we imagine a sample of random patent pairs constructed by pairing each of our potentially cited patents with a random draw of potentially citing patents. We could model the likelihood of a patent citation in this sample as a Bernoulli outcome y that equals 1 with a probability

$$\Pr(y = 1 \mid x = x_i) = \Lambda(x_i\beta) = \frac{1}{1 + e^{-x_i\beta}}.$$

Here, i is an index for the sample of potential citations (i.e., patent pairs), $\mathbf{x}_i$ represents the vector of covariates and controls (described later), and $\boldsymbol{\beta}$ is the vector of parameters to be estimated. Because the likelihood of a focal patent being cited by a random patent is extremely small, it is not practical to carry out the estimation based solely on a data set constructed via random sampling of possible pairs of patents. Instead, we employ a "choice-based" sample, wherein the sampled fraction γ of potentially citing patents that actually cite a focal patent is much larger than the fraction α of the patents not involved in a real citation to it. The usual (unweighted) logistic estimation based on such a sample would be biased, because the sampling rate is different for different values of the dependent variable. One way to avoid the bias is the *weighted exogenous sampling maximum likelihood* (WESML) approach, which involves a modified logistic estimation based on weighting each observation by the reciprocal of the ex ante probability of its inclusion in the sample (Manski and Lerman 1977).[8]

The basic WESML approach as previously described is based on employing a sample where the "zeroes" are drawn from the population of unrealized citations with the same ex ante likelihood. Recognizing that technological relatedness is a particularly strong driver of citation likelihood between patents, we can refine the choice-based sampling approach further to also get benefits from stratification on this explanatory variable. This implies allowing the parameter α to vary across different $y = 0$ subpopulations (Manski and McFadden 1981; Amemiya 1985, Chap. 9). Indeed, by carefully considering the respective subpopulations (defined by different technology classes and years of origin) from which we have effectively drawn our JTH-style control patents in the previous section, we can interpret our matched sample as above and appropriately calculate the weights to use with each control pair. However, as the appendix explains in more detail, this is not sufficient in itself. Using the WESML approach with the matched sample also requires extending the sample to ensure representation

[8] See the appendix for further details. See also Greene (2003, Chap. 21) for a discussion of choice-base sampling.

of potentially citing patents belonging to years and/or technology classes not represented in the original patent citations (and hence in the resulting matched sample). Doing this ensures that the strata considered are not only mutually exclusive but also exhaustive in representing the full population of potential citations. The above steps lead to our final sample of 13,728,582 patent pairs, which includes 4,007,217 actual citations (taking $\gamma = 1$), 4,007,217 JTH-style matched pairs, and 5,714,148 additional pairs from citing classes and years not represented in the matched sample. An example included in the appendix further illustrates the above sampling procedure as well as calculation of appropriate weights for all the control observations.

Rather than making specific functional form assumptions about temporal patterns, we account for citation lag (i.e., years elapsed between the cited and citing patents in a pair) nonparametrically by employing a full set of indicator variables. This is in addition to accounting for the cited patent's time period of origin using a separate set of indicator variables (Rysman and Simcoe 2008). Relying on longitudinal variation in our sample, we are therefore able to separately identify cohort effects and citation lag effects in a way that previous studies with more restrictive samples (such as Thompson 2006) were not able to. We also include indicators for the cited patent's two-digit National Bureau of Economic Research (NBER) technological subcategory.[9] Finally, because the citation probability might also be driven by other characteristics of the cited patent, we control for observables and cluster standard errors to account for unobserved ones.

3. Extending the Traditional Matching Approach

Before turning to our regression approach in the next section, we present some analysis that extends the more traditional JTH-style approach. This should allow a reader familiar with prior literature to better relate our study to existing research in terms of both what kind of findings remain similar across the two approaches and which new insights emerge specifically from using the regression approach.

3.1. Baseline Analysis Comparable with Prior Work

Following the empirical approach of Jaffe et al. (1993), we compare the incidence of geographic collocation of the potential knowledge sources as represented in actual citations as well as matched control pairs, in turn using the country, state, and metropolitan area as the geographic units of analysis. As the side-by-side comparison in Table 1 shows, our findings at each of the three units of analysis are quite comparable to those reported by Jaffe et al. (1993) as well as to a replication by Thompson and Fox-Kean (2005). The incidence of collocation for all three geographic units is statistically and economically greater between actual citations and the corresponding matched control pairs: 74.7% versus 57.6% at the country level; 13.4% versus 6.2% at the state level; and 7.6% versus 2.6% at the metropolitan level.[10]

3.2. Further Investigation of the Border Effects Using the Traditional Matching Framework

It is difficult within the JTH framework to separate the extent to which localization spillovers are driven primarily by political borders, spatial proximity, or both. However, we can carry out at least some informative analysis. In doing so, we focus most on the robustness and nature of the state border effect because, although localization at the country level might be less surprising, the presence of a localization effect truly associated with state borders within a country like the United States is puzzling.

A first step in separating border and proximity effects even within the JTH framework is checking whether the state finding is driven by observations geographically distant from the state border. Columns (1)–(4) in Table 2 report findings from a JTH-style analysis using a subsample where the distance of a potentially cited patent's originating town or city to the closest state border is not more than 20 miles. If state borders played no role in knowledge diffusion and prior findings were driven entirely by observations distant from the borders, the state result ought to now disappear. Comparing column (2) in Table 1 with column (2) in Table 2, we find that does not happen. Even though state-level collocation in column (2) is substantially lower for citations in the near-border sample than the whole sample (7.1% in Table 2 versus 13.4% in Table 1), the matched pair sample collocation incidence is also lower (2.7% in Table 2 versus 6.2% in Table 1) so the ratio reported calculated in column (4) is in fact higher in Table 2. In other words, taking account of geographic distribution of technological activity, we find no evidence that not accounting for distance from a state border is somehow driving the state effect reported earlier.

Although the above analysis based on a subset of *cited* patents (representing the source of knowledge)

[9] The NBER classification we refer to is drawn from Appendix 1 in Jaffe and Trajtenberg (2002, Chap. 13).

[10] When Thompson and Fox-Kean (2005) subsequently employ nine-digit technology matching, they find that over two-thirds of their patents cannot be matched. Our approach is instead to stick to a three-digit initial match but control for a finer technological level through additional variables introduced directly into our regression model.

Table 1 **Replicating Findings from Previous Studies**

	Our matched sample				Jaffe et al. (1993) sample				Thompson and Fox-Kean (2005) three-digit sample			
	(1) Citations sample	(2) Intraregion citations (%)	(3) Intraregion controls (%)	(4) Ratio (2)/(3)	(5) Citations sample	(6) Intraregion citations (%)	(7) Intraregion controls (%)	(8) Ratio (6)/(7)	(9) Citations sample	(10) Intraregion citations (%)	(11) Intraregion controls (%)	(12) Ratio (10)/(11)
Country-level analysis	4,007,217	74.7	57.6	1.30	7,759	68.0	61.4	1.11	7,627	68.6	55.6	1.23
State-level analysis	4,007,217	13.4	6.2	2.16	7,759	9.7	5.1	1.90	7,627	7.8	5.0	1.55
Metropolitan-level analysis	4,007,217	7.6	2.6	2.92	7,759	6.6	1.7	3.88	7,627	5.2	3.5	1.50

Notes. The Jaffe et al. (1993) numbers reported here were calculated based on pooling of results for their different subsamples primarily using information available in their Table 3 in a manner similar to that reported by Thompson and Fox-Kean (2005). The Thompson and Fox-Kean (2005) sample statistics are for the first sample they construct by employing three-digit technology matching to be comparable to that in Jaffe et al. (1993). Whereas Thompson and Fox-Kean (2005) subsequently construct other samples using more fine-grained technology matching, we instead rely on regression models to similarly account for technology more finely. Using formal t-tests confirmed that difference of means between incidences of geographic collocation for actual citations versus corresponding controls were statistically significant in all cases, so the t-statistics have not been reported to conserve space.

Table 2 **Further Investigation of the State Border Effect**

	Cited patent from near a state border				Cited patent from near a state border and citing patent from focal state dyad				Cited patent from near a state border and citing patent from focal state dyad as well as same CBSA as cited patent			
	(1) Citations sample	(2) Intraregion citations (%)	(3) Intraregion controls (%)	(4) Ratio (2)/(3)	(5) Citations sample	(6) Intraregion citations (%)	(7) Intraregion controls (%)	(8) Ratio (6)/(7)	(9) Citations sample	(10) Intraregion citations (%)	(11) Intraregion controls (%)	(12) Ratio (10)/(11)
Country-level analysis	996,627	74.9	58.4	1.28								
State-level analysis	996,627	7.1	2.7	2.63	93,703	68.4	55.7	1.23	40,784	87.2	82.8	1.05
Metropolitan-level analysis	996,627	6.1	2.2	2.77	93,703	55.8	38.0	1.47				

Notes. To ensure that within-state localization reported above is not just a distance effect driven by cited patents in a state's interior, columns (1)–(4) carry out the JTH-style analysis using a subsample of our matched sample where the distance of the cited patent's originating town or city to the closest state border is not more than 20 miles. In columns (5)–(8), the set of actual citations is restricted to those arising either within the cited patents or in the closest neighboring state—with the set of control citations to use as a benchmark also being regenerated based on a matching with all potentially citing patents within these two states using their application year and a three-digit technology class. In columns (9)–(12), as an additional robustness check to distinguish the effect of metropolitan collocation from state borders, analysis has been further restricted to cited patents originating in CBSAs that cross state borders and have both actual as well as corresponding control citations within the CBSA (with one or both of them potentially still crossing the state border). Difference of means between incidences of geographic collocation for actual citations versus corresponding controls were statistically significant in all cases, so the t-statistics have not been reported to conserve space.

originating near a state border increases confidence in the possibility that state borders do indeed have an independent effect, columns (5)–(8) refine this by restricting the set of potentially *citing* patents to those that originate within one of two states separated by the state border under consideration. For example, for a cited patent from Haverhill, Massachusetts (near the New Hampshire border), we would consider only (potential) citations from either Massachusetts or New Hampshire. Given that the citing patents in the matched pairs in our original sample could be from anywhere, this analysis relies on a new matched sample appropriate to the task. A control patent is now generated by matching the citing patent to a patent not just from the same three-digit technology class and the same year but also originating from within the state dyad being considered.

The interpretation of the results reported in columns (5)–(8) is that, in a sample comprising only dyads of neighboring states, knowledge generated within 20 miles of a state border is still much more likely to be used within its state of origin than in the neighboring state (after, as before, adjusting for geographic distribution of different technology classes). In other words, the finding of a state border remains qualitatively robust to using this alternate methodology.[11] Because we use a new sample that restricts actual and potential citations to be between neighboring states within the United States, note that country border effects have been filtered out (so country-level analysis is no longer carried out) and that the reported numbers are also not comparable with the findings from columns (1)–(4).

One interesting feature of U.S. geography is that 62 of the 943 CBSAs include more than one state. For example, the CBSA containing Cincinnati, Ohio, also extends into sections of Kentucky and Indiana. This allows us to test the border effect by examining whether in-state localization exists even for knowledge flows within such CBSAs. Specifically, columns (9)–(12) report the findings based on a subsample of the data in columns (5)–(8) where the observations only include cited patents originating in a multistate CBSA. The observations are further restricted to citations coming from within the CBSA that are also matched to control citations also within the same CBSA. By construction, metropolitan effects have therefore been filtered out (so CBSA-level analysis is no longer carried out). Difference of means between incidences of state-level collocation for actual citations and the corresponding controls remains statistically significant. Although their ratio is now much smaller, it should be noted that this is a very conservative test using a smaller, highly restrictive within-CBSA sample. Thus, just the fact that we find *any* state-border effect in this case is perhaps in itself quite remarkable. To a skeptic, this could be an indication instead that the state border effect is perhaps not as strong as it is made out to be in the earlier analysis. At this point, we are agnostic to an exact interpretation—preferring instead to address the issue using our regression framework.

3.3. Analyzing Long-Term Localization Trends Using the Traditional Matching Framework

Our sample size is orders of magnitude larger than those employed in previous studies, so we can carry out more detailed analyses as reported in Table 3. Columns (1)–(4) segment our cited patents drawn from 1975–2004 into six five-year periods.[12] Localization of knowledge spillovers remains robust across all periods for all three geographic units. Further, we can examine the time trends by taking the ratio of collocation frequency for inventor pairs comprising actual citations versus matched controls reported in column (4) as an indicator of the strength of the geographic effects. What is rather striking is that—despite much talk about globalization and decreasing relevance of geographic separation—the role of geography appears to have increased rather than decreased over time. Given that the JTH framework only analyzes each geographic unit in isolation, this analysis is, however, not able to disentangle whether the time trends are reflective primarily of underlying border effects, proximity effects, or a combination of the two. We will therefore return to this issue later in the context of our preferred regression framework that accounts for all geographic effects simultaneously.

3.4. Analyzing Inventor vs. Examiner Citations Using the Traditional Matching Framework

Recent work has noted that many patent citations are included not by the inventors themselves but later by patent examiners (Alcacer and Gittelman 2006). Therefore, it is useful to carry out an analysis complementary to the above by examining just inventor-added citations, because these might arguably be more likely to reflect prior art that an inventor was aware of in coming up with the focal

[11] In choosing the sample of cited patents near state borders, we have reported findings based on a cut-off of 20 miles as a compromise between being close to the border and having a reasonable sample size. We tried progressively smaller windows starting from 50 miles and going all the way down to those within 5 miles of a state border. The findings remained robust in support of a state border effect.

[12] The sample size drops during the last five-year period (2000–2004) because, although the earlier periods employ a full 10-year citation window, for this period we only observe citing patents through 2009.

Table 3 **Distinguishing Different Time Periods and Citations Added by Inventors vs. Examiners**

	Full matched sample				Inventor-added citation subsample				Examiner-added citation subsample			
	(1) Citations sample	(2) Intraregion citations (%)	(3) Intraregion controls (%)	(4) Ratio (2)/(3)	(5) Citations sample	(6) Intraregion citations (%)	(7) Intraregion controls (%)	(8) Ratio (6)/(7)	(9) Citations sample	(10) Intraregion citations (%)	(11) Intraregion controls (%)	(12) Ratio (10)/(11)
Country-level analysis												
1975–1979	262,657	66.7	59.0	1.13								
1979–1984	307,090	67.5	56.5	1.19								
1985–1989	504,546	73.4	58.0	1.27								
1990–1994	941,141	76.1	57.3	1.33	360,541	85.1	57.5	1.48	154,186	59.7	55.1	1.08
1995–1999	1,496,672	77.0	58.2	1.32	917,811	85.4	59.0	1.45	495,037	62.1	56.6	1.10
2000–2004	495,111	75.0	55.9	1.34	288,992	85.4	57.1	1.50	203,926	60.3	54.2	1.11
State-level analysis												
1975–1979	262,657	8.9	4.6	1.93								
1979–1984	307,090	9.4	4.5	2.09								
1985–1989	504,546	11.1	4.9	2.27								
1990–1994	941,141	13.4	5.8	2.31	360,541	15.7	6.1	2.57	154,186	9.5	5.6	1.70
1995–1999	1,496,672	14.7	7.1	2.07	917,811	16.8	7.2	2.33	495,037	10.8	6.9	1.57
2000–2004	495,111	16.3	7.3	2.23	288,992	19.3	7.5	2.57	203,926	12.1	7.1	1.70
Metropolitan-level analysis												
1975–1979	262,657	5.3	2.1	2.52								
1979–1984	307,090	5.6	2.1	2.67								
1985–1989	504,546	6.7	2.1	3.19								
1990–1994	941,141	8.0	2.5	3.20	360,541	9.4	2.6	3.62	154,186	5.3	2.3	2.30
1995–1999	1,496,672	7.9	2.8	2.82	917,811	9.1	2.9	3.14	495,037	5.6	2.7	2.07
2000–2004	495,111	9.4	2.9	3.24	288,992	11.4	3.0	3.80	203,926	6.7	2.7	2.48

Notes. Columns (1)–(4) employ exactly the same matched sample as the corresponding columns in the previous table except that the analysis has now been broken up into six five-year time periods based on the application year of the cited patent. The sample size drops during 2000–2004 because, although the first five periods employ the full 10-year citation window, the observed window is shorter for patents in this period, given that we only observe citing patents until 2009. Columns (5)–(8) are based only on the subsample of citations added by inventors and their corresponding controls, and columns (9)–(12) are based only on the subsample of citations added by examiners and their corresponding controls. Because this distinction is only available for citing patents post-2001, this analysis is done only for the cited patent originating periods for which the citation window overlaps with availability of the inventor versus examiner distinction information for citations.

invention.[13] Columns (5)–(8) of Table 3 report the JTH-kind analysis based only on the subsample of citations added by inventors (and the corresponding controls). Because the inventor/examiner distinction is only available for citations post-2001, these calculations are reported only for the cited patent originating during one of the three five-year periods for which the citation window overlaps with availability of this information for a significant fraction of the citations. Comparing the extent of the localization effect calculated in column (8) with column (4) reveals that a focus on just inventor-added citations significantly strengthens the geographic localization for all three geographic units of analysis. Unlike the results in column (4), the results in column (8) do not show any time trends—though that is largely reflective of the fact that the analysis cannot even be carried out for the first three periods because of unavailability of the inventor versus examiner distinction for citing patents pre-2001.

Thompson (2006) exploits the inventor/examiner distinction to address a challenge when using a JTH-style matching approach: because even the finest available technological classification might not capture some unobserved technological characteristics driving both patent citation patterns and geographical collocation, it is hard to be definitive about geographic collocation *leading to* increased knowledge diffusion. He suggests an identification strategy wherein one only takes greater localization for inventor-contributed citations *relative to* that for citations added by examiners as reliable evidence of localized knowledge spillovers. The rationale is twofold. First, because patent examiners are generally recruited directly after college, they do not have any specialized work experience that could bias them toward adding citations to patents from specific locations. Second, these examiners work at a single campus in Alexandria, Virginia, further making them "geography blind." In other words, the examiners should have no reason to disproportionately add localized citations over and above the natural distribution of prior patents relevant for a given patent application, making examiner-added citations useful as an appropriate benchmark for interpreting geographic localization of inventor-added citations. To use this approach for analyzing inventor citation findings from columns (5)–(8), we first report analysis using just examiner-added citations (and corresponding matched controls) in columns (9)–(12). Comparing columns (8) and (12), the calculated ratio between collocation incidences for realized citations versus matched patent pairs is found to be higher in all cases for inventor-added citations than for examiner-added citations, further establishing the robustness of the finding on geographic localization of knowledge spillovers. Note that we still have not disentangled border versus proximity effects, for which we turn to their simultaneous examination in our regression framework.

[13] In addition to the inventor versus examiner distinction being available only post-2001, a case can also be made in favor of considering all citations (rather than just inventor citations) because inventors may omit—even deliberately, for strategic reasons—reference to some patents representing knowledge they build on (Lampe 2012).

4. Analysis Using the WESML Regression Framework

4.1. Simultaneous Examination of Multiple Geographic Levels

We now turn to the regression framework to simultaneously examine national and state borders after accounting for proximity effects related to metropolitan (i.e., CBSA) collocation and geographic distance. Table 4 summarizes the variables used in our analyses. It is worth restating the data limitation that *distance* is measured based on the latitude and longitude coordinates we have for inventor cities, not the exact inventor addresses. Before trying to disentangle borders and proximity, however, it is instructive to get an overall sense of diffusion and geography. The analysis reported in column (1) of Table 5 is the simplest way of seeing this. The WESML regression estimates have an intuitive interpretation in terms of how an explanatory variable drives the likelihood of citation between random patents in the population, with the fact that citations are rare events making it possible to in fact directly interpret the logistic model coefficients as percentage effects on citation likelihood.[14] Column (1) implies that the likelihood of citation falls by 36% with a doubling of distance.

The analysis reported in column (2) also includes relevant control variables. This includes controls for technological similarity and relatedness between patents using a series of associated variables rather than only relying on the three-digit technology class match. The findings in column (2) imply that the likelihood of citation now falls by 27% with a doubling

[14] In a logistic model, the marginal effect for a variable j is $\beta_j \Lambda'(\mathbf{x\beta})$, which turns out to equal $\beta_j \Lambda(\mathbf{x\beta})[1-\Lambda(\mathbf{x\beta})]$. In general, this would need to be calculated either based on the mean predicted probability or using the sample mean for $\Lambda(\mathbf{x\beta})$. But the fact that citations are rare events allows further simplification: since $\Lambda(\mathbf{x\beta})$ is much smaller than 1, $\beta_j \Lambda(\mathbf{x\beta})[1-\Lambda(\mathbf{x\beta})]$ is practically equivalent to $\beta_j \Lambda(\mathbf{x\beta})$. This means the coefficient estimate for β_j can be directly interpreted as the percentage change in citation probability with a unit change in variable j.

Table 4 Definitions of Variables Used During Regression Analysis

Political border variables	
same country	Indicator variable that is equal to 1 if the citing and cited patents originate in the same country, that is, the United States (given that our cited patent sample is drawn from the United States only)
same state	Indicator variable that is equal to 1 if the two patents originate in the same state (within the United States)
Spatial proximity variables	
same cbsa	Indicator variable that is equal to 1 if the citing and cited patents originate from inventors located in the same core based statistical area (CBSA) as per the 2003 definition of CBSAs by the U.S. Office of Management and Budget (CBSA definitions are meant to cover reasonable commuting distances and replace the prior MSA/PMSA/CMSA definitions for defining U.S. metropolitan areas in a more standardized fashion)
distance	Distance, in miles, between the cities where the first inventors of the source and destination patents live (calculated as spherical distance between the latitude and longitude values for these cities)
Technological relatedness variables	
same tech category	Indicator variable that is equal to 1 if the two patents belong to the same one-digit NBER technology category
same tech subcategory	Indicator variable that is equal to 1 if the two patents belong to the same two-digit NBER technical subcategory
same tech class	Indicator variable that is equal to 1 if the two patents belong to the same three-digit USPTO primary technology class
relatedness of tech classes	Likelihood of citation (scaled by 100) between random patents with the same respective three-digit primary technology classes that the focal cited and citing patents belong to
overlap of tech subclasses	Natural logarithm of one plus the number of overlapping nine-digit technology subclasses under which the patents are categorized
Patent-level variables	
references to other patents	Number of references the cited patent makes to other patents
references to nonpatent materials	Number of references the cited patent makes to published materials other than patents
number of claims	Number of claims the cited patent makes
period	A sequential number representing which of our six five-year time periods the focal cited patent belongs to: 197,579 being period 0, 1980–1984 being period 1, 1985–1989 being period 2, 1990–1994 being period 3, 1995–1999 being period 4, and 2000–2004 being period 5

of distance—the difference between the citation likelihoods corresponding to column (1) and (2) estimates being driven mainly by the new technology controls. In line with the JTH argument, we find knowledge flows within the same or related technologies to be stronger than those across different technologies, as indicated by the positive and significant estimates for *same tech category*, *same tech subcategory*, *same tech class*, and *relatedness of tech classes*. Taking into account the Thompson and Fox-Kean (2005) critique regarding the inadequacy of three-digit technological controls, we have also included a control variable to capture overlap between the citing and cited patent along their secondary nine-digit technology subclasses (*overlap of tech subclasses*); we find that to have a strong effect as well.

Setting the geographic distance variable aside for now, columns (3)–(5) successively introduce variables for collocation at three geographic levels: country (*same country*), state (*same state*), and metropolitan area (*same cbsa*). In terms of magnitude, column (5) estimates imply a 77% greater likelihood of within-country knowledge flow than across national borders, 41% greater likelihood for within-state flow than that across state borders, and 77% greater likelihood for within-CBSA flow than across CBSA boundaries. The effect size corresponding to each of these three estimates is smaller than what it would be if estimating these effects individually in separate models, and nonoverlap of the corresponding confidence intervals is indicative of this difference being statistically significant. This highlights the benefit of using a regression approach in disentangling effects at the various geographic levels by simultaneously considering the effect of all three.[15]

Simultaneously considering multiple geographic units indicates that there is more to the national and state border effects than a mere aggregation of localization mechanisms operating at the metropolitan level. The estimates in column (5), however, do not rule out the possibility that such effects are not epiphenomenal with spatial distance, because including the CBSA collocation variable does not account for distance-related effects that might be more gradual than CBSA collocation. To this point, the model in column (6) now also includes the distance variable from before. As expected, with geographic proximity now better controlled for through the combination of metropolitan collocation and distance, both border effects become smaller. The extent of this drop—calculated in terms of a percentage difference in the

[15] If we carry out this analysis excluding the nine-digit technology control, the magnitude of geographic localization on all three dimensions turns out to be larger—with the difference being the greatest for metropolitan collocation. This is in line with intuition that geographic concentration of technological activity—which is what our technology-related control variables account for—is greater when viewed at a finer level of granularity for technology.

Table 5 Simultaneous Consideration of Political Borders and Spatial Proximity

	(1) Full sample	(2) Full sample	(3) Full sample	(4) Full sample	(5) Full sample	(6) Full sample	(7) Full sample	(8) Near-border sample	(9) Excluding California	(10) Excl. comp. and comm.
same country			0.863***	0.769***	0.766***	0.535***	0.451***	0.513***	0.447***	0.441***
			(0.006)	(0.006)	(0.006)	(0.011)	(0.016)	(0.032)	(0.021)	(0.024)
same state				0.750***	0.405***	0.109***	0.228***	0.253***	0.346***	0.230***
				(0.017)	(0.024)	(0.027)	(0.027)	(0.047)	(0.049)	(0.044)
same cbsa					0.769***	0.456***	0.337***	0.295***	0.433***	0.407***
					(0.030)	(0.029)	(0.032)	(0.071)	(0.054)	(0.045)
ln(*distance* + 1)	−0.364***	−0.271***				−0.137***				
	(0.001)	(0.003)				(0.005)				
distance 0 (*i.e., same city*)							1.665***	1.896***	1.661***	1.910***
							(0.067)	(0.207)	(0.103)	(0.095)
distance 0–10 *miles*							1.129***	1.342***	1.203***	1.236***
							(0.057)	(0.112)	(0.083)	(0.086)
distance 10–20 *miles*							0.990***	1.020***	0.923***	1.064***
							(0.049)	(0.088)	(0.073)	(0.070)
distance 20–30 *miles*							0.836***	0.517***	0.720***	0.919***
							(0.055)	(0.108)	(0.081)	(0.079)
distance 30–40 *miles*							0.552***	0.239*	0.386***	0.660***
							(0.071)	(0.125)	(0.116)	(0.100)
distance 40–50 *miles*							0.613***	0.433***	0.657***	0.802***
							(0.099)	(0.096)	(0.071)	(0.067)
distance 50–75 *miles*							0.595***	0.533***	0.583***	0.707***
							(0.040)	(0.066)	(0.050)	(0.052)
distance 75–100 *miles*							0.546***	0.529***	0.579***	0.665***
							(0.038)	(0.066)	(0.049)	(0.053)
distance 100–150 *miles*							0.599***	0.584***	0.614***	0.656***
							(0.033)	(0.053)	(0.040)	(0.048)
distance 150–200 *miles*							0.585***	0.553***	0.584***	0.670***
							(0.029)	(0.050)	(0.033)	(0.039)
distance 200–300 *miles*							0.479***	0.557***	0.491***	0.544***
							(0.029)	(0.043)	(0.035)	(0.042)
distance 300–400 *miles*							0.503***	0.520***	0.566***	0.567***
							(0.024)	(0.044)	(0.027)	(0.034)
distance 400–500 *miles*							0.479***	0.569***	0.508***	0.566***
							(0.024)	(0.048)	(0.029)	(0.036)
distance 500–750 *miles*							0.480***	0.494***	0.473***	0.519***
							(0.022)	(0.038)	(0.027)	(0.033)
distance 750–1,000 *miles*							0.439***	0.483***	0.450***	0.484***
							(0.020)	(0.038)	(0.025)	(0.029)
distance 1,000–1,500 *miles*							0.419***	0.433***	0.441***	0.467***
							(0.020)	(0.038)	(0.025)	(0.030)
distance 1,500–2,000 *miles*							0.377***	0.400***	0.391***	0.405***
							(0.019)	(0.040)	(0.024)	(0.027)
distance 2,000–2,500 *miles*							0.368***	0.470***	0.417***	0.382***
							(0.019)	(0.038)	(0.025)	(0.028)
distance 2,500–4,000 *miles*							0.461***	0.487***	0.460***	0.507***
							(0.015)	(0.023)	(0.016)	(0.021)
distance 4,000–6,000 *miles*							0.112***	0.257***	0.154***	0.133***
							(0.010)	(0.027)	(0.011)	(0.012)
same tech category		1.103***	1.115***	1.111***	1.108***	1.106***	1.107***	1.088***	1.102***	0.893***
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.011)	(0.006)	(0.007)
same tech subcategory		1.298***	1.310***	1.300***	1.299***	1.297***	1.296***	1.298***	1.300***	1.460***
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.015)	(0.009)	(0.010)
same tech class		2.141***	2.154***	2.156***	2.145***	2.144***	2.144***	2.267***	2.215***	2.283***
		(0.016)	(0.014)	(0.016)	(0.015)	(0.015)	(0.014)	(0.025)	(0.017)	(0.016)

Table 5 (Continued)

	(1) Full sample	(2) Full sample	(3) Full sample	(4) Full sample	(5) Full sample	(6) Full sample	(7) Full sample	(8) Near-border sample	(9) Excluding California	(10) Excl. comp. and comm.
relatedness of tech classes		1.512***	1.604***	1.481***	1.518***	1.501***	1.502***	1.537***	1.650***	1.567***
		(0.129)	(0.104)	(0.127)	(0.112)	(0.116)	(0.104)	(0.193)	(0.138)	(0.124)
overlap of tech subclasses		1.687***	1.691***	1.686***	1.686***	1.684***	1.681***	1.716***	1.704***	1.845***
		(0.011)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.019)	(0.013)	(0.016)
ln(*references to other patents* + 1)		0.134***	0.135***	0.135***	0.136***	0.135***	0.135***	0.159***	0.149***	0.134***
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.012)	(0.006)	(0.007)
ln(*references to nonpatent materials* + 1)		0.034***	0.034***	0.034***	0.033***	0.033***	0.033***	0.007	0.032***	0.039***
		(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.008)	(0.006)	(0.008)
ln(*number of claims*)		0.092***	0.092***	0.092***	0.093***	0.092***	0.092***	0.110***	0.095***	0.082***
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.010)	(0.006)	(0.006)
Period indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citation lag indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-digit tech indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	3,600,000	10,994,852	10,474,569
Pseudo-R^2	0.0122	0.181	0.179	0.181	0.182	0.182	0.183	0.189	0.188	0.196
Wald chi^2	124,220	756,991	785,451	767,431	759,980	755,320	763,511	221,768	616,233	519,519
Degrees of freedom	1	69	69	70	71	72	91	91	90	87

Notes. The unit of observation is a pair of patents representing an actual or potential citation. The dependent variable is an indicator for whether or not the potentially citing patent actually cited the focal patent. A choice-based stratified sample is used, and a weighted logistic regression (WESML) approach is implemented using observation weights that reflect sampling frequency associated with different strata. The regression model also uses a constant term and indicator variables as indicated above, but these are not reported to conserved space and are available from the authors on request. Robust standard errors are shown in parentheses and are clustered on the cited patent.

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

average predicted marginal effect for a variable across the two models—turns out to be much larger for the *same state* effect than the *same country* effect.

To allow more flexibility in how distance constrains knowledge flows, column (7) repeats the analysis with a series of indicator variables for distance ranges, covering increasing distances starting in the sequence *distance* 0 *miles* (i.e., same city), *distance* 0–10 *miles*, and so on. The omitted category is distance greater than 6,000 miles. This nonparametric approach ensures that the *same country* and *same state* estimates more accurately measure border effects independent of geographic proximity. (Even more fine-grained indicators did not materially alter findings.) Not surprisingly, estimates for the distance indicators themselves reveal that knowledge flows are greatest when the source and recipients are collocated within the same city (i.e., distance = 0) and that the distance effect gradually falls (more or less monotonically) with distance. Once more, however, we find statistically and economically significant estimates for *same country* and *same state* even after we have accounted for geographic proximity using *same cbsa* and distance indicators.[16] (Note that it is hard to directly compare *same country* and *same state* coefficients across columns (6) and (7), as the latter has a large number of new variables in the form of distance indicators.) This finding challenges an interpretation that localized knowledge diffusion reported by previous studies is merely a manifestation of intraregional distances being on average smaller than cross-regional distances.[17]

4.2. Further Investigation of the Border Effects Using the WESML Regression Framework

We now examine subsamples to figure out whether our findings are driven by particular kinds of patents. One concern might be whether the state-level finding is driven by observations that are quite distant from the state border. Analogous to the near-border analysis presented for the JTH approach, we analyze diffusion of knowledge originating near state borders to see if there is on average a similar state-border effect even for these. Specifically, we look at the subset of potentially cited patents that lie within 20 miles

[16] Rather than calculating standard errors based on clustering at the patent level, we also tried the associate editor's suggestion of geographic clustering at different levels—the city, the CBSA, and even the state—to be conservative in the kinds of unobserved heterogeneity accounted for. Although the standard errors did become larger as expected, the coefficients for *same country, same state*, and *same cbsa* still remained statistically significant at the 1% level.

[17] In additional analysis, we tried models with indicators for *contiguous countries* and *contiguous states* to distinguish cases where the source and destination share a border. While we did find knowledge flow to be more intense between contiguous regions, we found that independent country and state border effects persist.

of a state border. As column (8) in Table 5 indicates, the findings for the near-border cited patent subsample turn out to be qualitatively similar to those from the full sample (column (7)), including the continued presence of a significant same-state effect.

Next, we subset our sample by removing California, as Silicon Valley has been often described as an outlier for diffusion (Almeida and Kogut 1999). As the top state in terms of patenting activity and one of the largest in terms of area, one might worry that our results depend on California in ways that state fixed effects do not capture. In column (9) of Table 5, both country and state localization are found to be robust to excluding California. To further investigate whether our findings are state specific, in an analysis not reported to conserve space, we also carried out analogous analyses for cited patent subsamples from the 10 largest patenting states. The findings revealed that, in 6 of these 10 cases, observed state-level localization of knowledge originating within the state borders could not be completely explained simply by geographic proximity effects in the form of metropolitan collocation and/or shorter geographic distances. In other words, the finding is not driven by just one or two specific states. In fact, California turned out to be one of the minority cases where state borders *do not* seem to have an effect independent of distance (but CBSA boundaries like those of the Silicon Valley still *do*), suggesting that—once one crosses out of certain areas like Silicon Valley—knowledge is no longer further constrained by state borders over and above effects related simply to distance.

Next, we turn to checking whether the results could similarly be driven by specific sectors. To start with, we exclude the one-digit NBER technology category *computers and communications*—a sector many scholars consider unique. As column (10) in Table 5 shows, the results are qualitatively unchanged. We also carried out (but omit for space) separate analyses for cited patent subsamples from all six one-digit NBER categories. We found that the findings are not driven by a specific sector. In fact, in five of the six cases, observed state-level localization of knowledge could not be completely explained simply by geographic proximity effects, the only exception being the NBER category *others*. Similarly, repeating the analysis with two-digit NBER subcategories revealed an independent state border effect for 30 of the 36 subsamples. Thus, the state border finding is clearly not driven by just one or two specific sectors either.

4.3. Analyzing Long-Term Localization Trends Using the WESML Regression Framework

We next investigate whether these effects are driven by particular time periods as opposed to being persistent. Before disentangling long-term trends in border and proximity effects, it is useful to start with an overall sense of how the role of geography in knowledge diffusion has evolved over time. With this view, column (1) in Table 6 extends the analysis from column (2) in Table 5 by adding an interaction term $period * \ln(distance + 1)$ between the distance variable and the time-period variable capturing the five-year period when the cited patent originated. (See Table 4 for detailed definition.)[18] Surprisingly, and contrary to the widespread notion that the importance of distance has been eroding over time because of globalization and technological advancement, the decay in citation rate with distance seems to have *increased* over time, albeit the economical magnitude of this is not too large.

In column (2), we turn to disentangling time trends in the border versus proximity effects, with the goal of figuring out whether the role of political borders has strengthened or weakened over time once proximity is accounted for. In addition to the distance variable, we re-introduce our other three geographic variables—*same country*, *same state*, and *same cbsa*—but now also bring in their interaction effects with the time variable *period*. The trends turn out to differ across different variables: the effect of national borders seems to have increased over time, whereas that for state borders and CBSA boundaries has decreased. Additional analyses in columns (3) and (4) add distance indicators and the full set of distance-period indicators, respectively, to more completely account for any distance-related effects and trends not captured above. The finding on the opposite time trends for country versus state borders remains qualitatively robust, with the country effect still strengthening over time and the state effect weakening. However, the CBSA finding is more fickle, becoming statistically insignificant in column (3) and ultimately flipping to become positive (and statistically significant) in column (4). This might be caused by the high correlation between the distance indicators and *same cbsa*.

As the model with the least functional form restrictions on distance, column (4) represents our specification of choice. Following Greene (2010), we interpret the results for the interaction terms in this nonlinear model graphically by calculating the average predicted effect of a 0 to 1 transition for *same country*, *same state*, and *same cbsa*. Specifically, by carrying out this exercise for the subsamples from different time periods, we plot the predicted effects for different periods in Figure 2. Examining the ratio of the predicted effect for the case where a specific

[18] Recall that we use separate sets of indicators for the cohort the cited patent comes from *and* the lag between the cited and citing patents. Using longitudinal variation, our sample allows us to separately identify cohort effects and lag effects in a way that previous studies with more restrictive samples (such as Thompson 2006) are not able to.

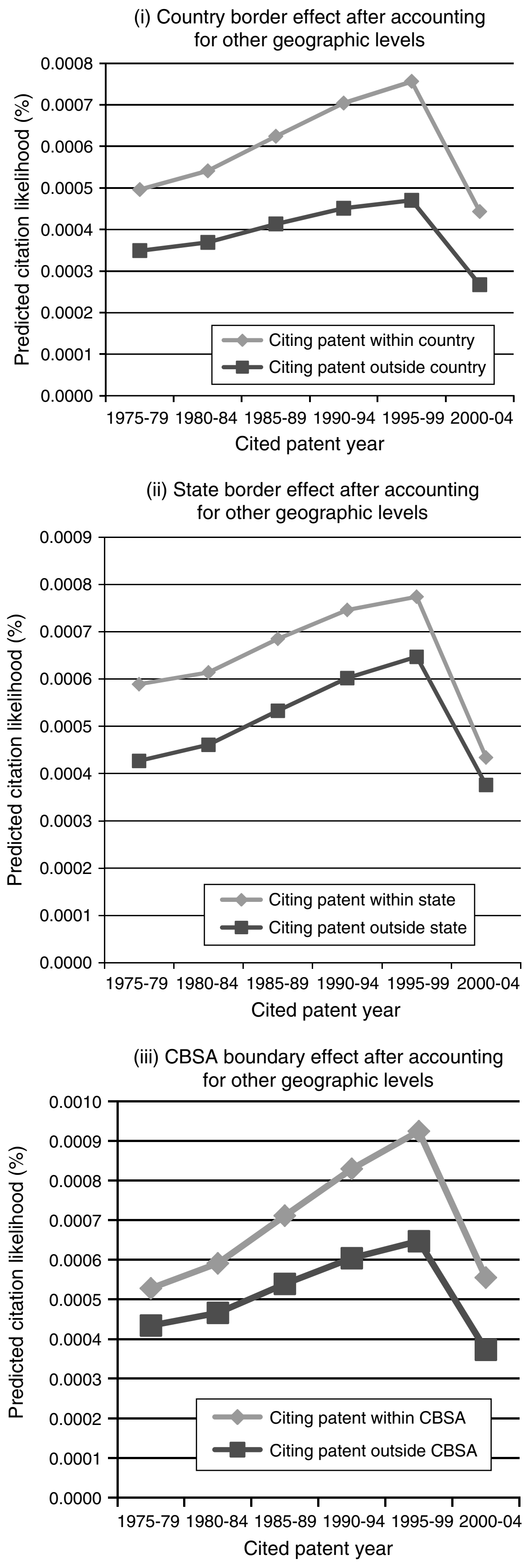
Table 6 Time Trends in Geographic Knowledge Diffusion Patterns

	(1) Full sample	(2) Full sample	(3) Full sample	(4) Full sample	(5) Full sample
same country		0.259***	0.140***	0.362***	0.249***
		(0.022)	(0.020)	(0.048)	(0.034)
same state		0.381***	0.367***	0.358***	0.332***
		(0.070)	(0.066)	(0.074)	(0.048)
same cbsa		0.616***	0.424***	0.125	−0.034
		(0.069)	(0.072)	(0.087)	(0.084)
ln(*distance* + 1)	−0.224***	−0.085***			
	(0.008)	(0.011)			
period ∗ *same country*		0.089***	0.106***	0.021*	
		(0.007)	(0.004)	(0.011)	
period ∗ *same state*		−0.089***	−0.048***	−0.035**	
		(0.017)	(0.016)	(0.015)	
period ∗ *same cbsa*		−0.053***	−0.026	0.048**	
		(0.019)	(0.020)	(0.020)	
period ∗ ln(*distance* + 1)	−0.011***	−0.017***			
	(0.002)	(0.003)			
same country ∗ *period 1980–1984*					0.061
					(0.080)
same country ∗ *period 1985–1989*					0.321***
					(0.051)
same country ∗ *period 1990–1994*					0.255***
					(0.047)
same country ∗ *period 1995–1999*					0.196***
					(0.043)
same country ∗ *period 2000–2004*					0.147***
					(0.053)
same state ∗ *period 1980–1984*					0.069
					(0.062)
same state ∗ *period 1985–1989*					−0.279**
					(0.133)
same state ∗ *period 1990–1994*					−0.011
					(0.060)
same state ∗ *period 1995–1999*					−0.171***
					(0.057)
same state ∗ *period 2000–2004*					−0.187***
					(0.072)
same cbsa ∗ *period 1980–1984*					0.113
					(0.114)
same cbsa ∗ *period 1985–1989*					0.513***
					(0.133)
same cbsa ∗ *period 1990–1994*					0.371***
					(0.100)
same cbsa ∗ *period 1995–1999*					0.412***
					(0.095)
same cbsa ∗ *period 2000–2004*					0.333***
					(0.119)
Distance-period indicators	No	No	No	Yes	Yes
Distance indicators	No	No	Yes	Yes	Yes
Period indicators	Yes	Yes	Yes	Yes	Yes
Citation lag indicators	Yes	Yes	Yes	Yes	Yes
Two-digit tech indicators	Yes	Yes	Yes	Yes	Yes
State indicators	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes
Number of observations	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582
Pseudo-R^2	0.181	0.183	0.183	0.183	0.183
Wald chi^2	758,828	754,616	762,307	778,608	780,425
Degrees of freedom	70	76	94	189	201

Notes. All notes from Table 5 apply here as well, except that regression coefficients for the control variables as well as for the distance-period and distance indicators (when applicable) are also omitted to further conserve space. As indicated, distance indicators are excluded in the first two models because a continuous distance variable has been directly included in those models.

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

Figure 2 Predicted Probabilities Across Different Time Periods



variable (such as *same country*) is set to 1 versus 0 helps interpret the economic magnitude of the trend. (Whether one of these individual predicted effect lines slopes upward or downward from a specific period to the next is not relevant to this analysis beyond how the ratio itself evolves over time.) For example, the ratio of the cases with *same country* being 1 versus 0 *increases* from 1.42 in 1975–1980 (predicted probabilities of 5.0 in a million for *same country* = 1 versus 3.5 in a million for *same country* = 0) to 1.66 in 2000–2004 (4.4 in a million versus 2.7 in a million). In contrast, the ratio of the cases with *same state* being 1 versus 0 *decreases* from 1.38 in 1975–1980 (probabilities of 5.9 in a million for *same state* = 1 versus 4.3 in a million *same state* = 0) all the way down to 1.15 in 2000–2004 (4.3 in a million versus 3.8 in a million). We offer a similar chart of CBSA for completeness; however, as our distance variable is based on a single latitude and longitude value for a city, we consider it too noisy to reliably disentangle microlevel distance effects from CBSA effects. We therefore suggest caution in interpreting the CBSA variables, treating these as controls for our study rather than taking the results as conclusive.

Relying on statistical testing using the *period* above has helped us formally test for long-term trends in knowledge diffusion patterns. However, for period-by-period findings that do not impose linear restrictions on the effect of *period*, column (5) depicts the results of interacting the geography variables, with indicators corresponding to different five-year time periods. The omitted (reference) period is 1975–1979. Relative to 1975–1979, the country border effects are stronger in four of the five subsequent periods (and statistically indistinguishable in the remaining one). In contrast, relative to the same baseline period, the state border effects are weaker in three of the five subsequent periods (and statistically indistinguishable in the two remaining ones). Overall, these results are generally consistent with the time trends documented above: country-level localization seems to have strengthened over time, whereas state-level localization has diminished.[19]

4.4. Analyzing Inventor vs. Examiner Citations Using the WESML Regression Framework

We now revisit the issue that many citations are generated not by inventors but by patent examiners. For easy interpretation, logistic regression estimates for the inventor versus examiner subsample

[19] We wondered about the extent to which the temporal patterns are an artifact of changes in sectoral composition over time. In analysis not detailed here, we found that the increase in country-level localization as well as the drop in state-level localization is a more general phenomenon than being driven by increasing dominance of specific sectors.

are first separately reported in columns (1)–(3) and columns (4)–(6), respectively, of Table 7. This is followed by the last three columns, which examine the two subsamples together in a single multinomial logistic framework in order to allow more rigorous inference. As noted in §3.4, unavailability of information on whether a pre-2001 citation was added by an inventor or examiner restricts our analysis to patents receiving a meaningful number of relatively recent citations. This reduces the number of observations considerably compared with Table 5 and also makes it impractical for exploiting the inventor/examiner citation distinction to shed further light on the long-term time trends.

We start with side-by-side analyses of the inventor-added citations subsample (which includes not only actual citations but also controls matched to those) in columns (1)–(3) and the examiner-added citations subsample (which also includes actual citations and corresponding controls) in columns (4)–(6) in Table 7. We compare columns (1) and (4) to assess the overall geographic effect. When not simultaneously accounting for political borders, the effect size implied by the coefficient on $\ln(\mathit{distance} = 1)$ variable is almost twice as large for citations added by inventors instead of by examiners. However, simultaneously considering all our geographic units representing political borders and spatial proximity in the remaining columns questions whether a large part of the overall effect is truly comprised of an impact of proximity per se. The difference between the effect sizes for coefficients on $\ln(\mathit{distance} + 1)$ for columns (2) and (5) are not that large, relative to the big gap between the two for columns (1) and (4). Similarly, the effects size for *same cbsa* is not too different between columns (2) and (5), and in fact becomes virtually indistinguishable between columns (3) and (6), as distance is accounted for nonparametrically in the form of our full set of indicator variables. This reinforces the concerns expressed by Thompson and Fox-Kean (2005) and Thompson (2006) that knowledge spillovers reported in earlier studies might, to a significant extent, have been a manifestation of the USPTO classification system (or, for that matter, *any* formal classification system), only imperfectly capturing true technological relationships across patents.

The previous finding on the influence of political borders, however, is not diluted as much by the inventor/examiner distinction. Comparing the estimates for *same country* in columns (2) and (3) with those in columns (5) and (6), respectively, examiner-added citations in fact show no country-level localization, whereas the effect for inventor citations is economically and statistically highly significant. The state-level result also remains robust. However, although the *same state* coefficient is statistically insignificant for the examiner-added citation analysis in column (5), it turns significant for the preferred specification in column (6) once distance is accounted for in a nonparametric fashion. Still, the effect size corresponding to the coefficient on *same state* in column (3) remains considerably larger than that in column (6). The relative weakness of the *same state* effect in this analysis might in part be a result of the limited timeframe of the inventor versus examiner distinction, as we are able to observe patents only in the latter portion of our 30-year window. Recall from the earlier time trends analysis that the same-state effect was weaker during this time period when considering all citations together. If we did have the inventor/examiner distinction data available for the earlier part of our sample, it is conceivable that the state border effect might have been stronger.

Directly comparing estimates from nonlinear regressions employing different subsamples (inventor versus examiner citations) relies on our earlier observation that these estimates have a natural interpretation in percentage terms because citations are rare events. While intuitive, this approach leaves two open questions. First, given the different control groups for inventor and examiner subsamples, this direct comparison could be problematic. Second, it is not straightforward to test hypotheses regarding statistical distinguishability of estimates across different models. To address these concerns, we pool the two subsamples and run the analysis as a single (weighted) multinomial regression for the three mutually exclusive and exhaustive outcomes possible for any pair of random patents: *inventor citation*, *examiner citation*, and *no citation*.

The first set of multinomial logit analyses, reported in columns (7)–(9), take *no citation* as the omitted (reference) category. The findings seem qualitatively very similar to those from separate logistic analyses for the two subsamples described above. In particular, most of the distinction between the coefficients on $\ln(\mathit{distance}+1)$ between the *inventor citation* and *examiner citation* outcomes again disappears in going from column (7) to column (8). In contrast, the coefficients for *same country* and *same state* remain much stronger for the *inventor citation* outcome than for the *examiner citation* outcome even in columns (8) or (9). The only distinction from before is that even the *same cbsa* effect now seems significantly stronger for the *inventor citation* case than for the *examiner citation* case, although the magnitude of the *same cbsa* difference is still somewhat smaller than that for the *same state* effect and much smaller than for the *same country* effect in the preferred model (9).

To formally test hypotheses comparing *examiner citation* estimates with the *inventor citation* estimates, columns (10)–(12) replicate the same multinomial

Table 7 Inventor-Added vs. Examiner-Added Citations

	(1) Inventor sample (logit)	(2) Inventor sample (logit)	(3) Inventor sample (logit)	(4) Examiner sample (logit)	(5) Examiner sample (logit)	(6) Examiner sample (logit)	(7) Full sample (multinomial logit)	(8) Full sample (multinomial logit)	(9) Full sample (multinomial logit)	(10) Full sample (multinomial logit)	(11) Full sample (multinomial logit)	(12) Full sample (multinomial logit)
Inventor citation												
same country		1.223***	0.753***					1.222***	0.754***		1.246***	0.809***
		(0.015)	(0.021)					(0.014)	(0.020)		(0.013)	(0.020)
same state		0.086***	0.246***					0.083***	0.243***		0.080***	0.125***
		(0.028)	(0.028)					(0.026)	(0.026)		(0.025)	(0.028)
same cbsa		0.417***	0.334***					0.411***	0.328***		0.095***	0.089**
		(0.035)	(0.041)					(0.032)	(0.037)		(0.031)	(0.035)
ln(*distance* + 1)	−0.352***	−0.166***					−0.353***	−0.168***		−0.189***	−0.028***	
	(0.004)	(0.007)					(0.004)	(0.007)		(0.003)	(0.006)	
Examiner citation												
same country					−0.034	−0.059		−0.024	−0.055**			
					(0.027)	(0.038)		(0.015)	(0.023)			
same state					−0.015	0.124**		0.002	0.118***			
					(0.056)	(0.053)		(0.031)	(0.033)			
same cbsa					0.408***	0.328***		0.316***	0.239***			
					(0.077)	(0.077)		(0.044)	(0.049)			
ln(*distance* + 1)				−0.173***	−0.147***		−0.163***	−0.140***				
				(0.006)	(0.013)		(0.004)	(0.008)				
Reference group:	No citation	No citation	No citation	No citation	No citation	No citation	No citation	No citation	No citation	Examiner citation	Examiner citation	Examiner citation
Distance indicators	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lag fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-digit tech fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	4,651,156	4,651,156	4,651,156	3,377,722	3,377,722	3,377,722	5,828,778	5,828,778	5,828,778	5,828,778	5,828,778	5,828,778
Wald chi^2	254,439	262,810	267,886	230,040	232,457	233,936	466,765	507,039	511,038	466,765	507,039	511,038
Degrees of freedom	66	69	88	66	69	88	132	138	176	132	138	176

Notes. All notes from Table 5 apply here, except that regression coefficients for the distance indicators and control variables are not shown. The first six columns employ weighted logistic regressions as before, but with only inventor-added citations and corresponding controls included in columns (1)–(3) and only examiner-added citations and corresponding controls included in columns (4)–(6). The last six columns employ weighted multinomial logistic regressions based on the combined sample, with the regression specifications used for columns (7)–(9) differing from those in columns (10)–(12) only in the reference category used (as indicated). This table employs only data from citing year 2001 onward because inventor versus examiner distinction is not available for earlier years. Given the resulting citation window of at most 10 years, all cited patents originating pre-1991 get dropped.

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

logit specification as in columns (7)–(9) after now taking *examiner citation* as the omitted category. This obviates the need to compare coefficient values across inventor and examiner citations in specification (9), instead allowing direct inspection of the equivalent column (12) coefficient significance as a formal test of the geography-related effects for political borders after accounting for spatial proximity. All three effects—*same country, same state*, and *same cbsa*—are found to be statistically significant in column (12). Additional statistical tests reveal that the *same country* effect is indeed significantly larger than the *same state* and *same cbsa* effects, although the latter two are statistically indistinguishable from each other. Overall, our main qualitative finding—that there are independent border effects even after geographic proximity is accounted for—therefore continues to hold even with a careful inventor versus examiner citation distinction. We can be quite confident in concluding that the border effect finding is indeed quite robust and not just an artifact of either the geographic proximity of inventors not being accounted for or there being measurement issues related to geographic distribution of technological activity.

5. Discussion and Conclusion

The contribution of this study is that it employs a novel regression framework based on choice-based sampling to *simultaneously* consider the impact of different geopolitical levels to help disentangle border effects from geographic proximity effects. In addition to accounting for technological relatedness between the citing and cited patents at multiple levels of granularity, we employ an identification approach inspired by Thompson (2006) to address concerns about unobserved aspects of technological relatedness. A robust finding of our study is that, on average, country and state borders serve as constraints on knowledge diffusion *even after accounting for geographic proximity* in the form of metropolitan collocation and geographic distance. We document that the findings are robust to examining only near-border samples and also not driven by specific states or sectors. In fact, application of the alternate identification strategy using the inventor/examiner distinction in citations only strengthens this finding regarding an independent effect of borders.

The finding that national borders have a strong effect might not be too surprising. The literature already suggests several border-related variables that future research could consider for digging deeper, such as linguistic, cultural, political, and economic differences between countries. Indeed, in an analysis not reported here, we found knowledge flows from the United States to other English-speaking countries to be stronger even after accounting for geographic distance. A more general treatment of this issue in the form of gravity-type models employed in international economics would, however, require a sample where not just the citing but also the cited patents are drawn from multiple countries.

One might still worry whether country-level results are driven by patents originating in different countries systematically differing in their propensity to cite USPTO patents for reasons unrelated to true knowledge flows. However, previous work such as Jaffe and Trajtenberg (2002, Chap. 7) and Singh (2007)—although not examining multiple geographic levels—has used USPTO data to report country-level localization effects of magnitude comparable to the United States in a number of other countries as well, suggesting that biases arising from systematic differences in propensity to cite are unlikely to be too large. Our additional finding that the country-level localization effect is driven almost entirely by citations that inventors themselves add also points to systematic differences in propensity to cite not being evident, at least in citations that USPTO examiners add later.

What is most surprising about our country-level finding is that it has only grown stronger over a period that has seen the rise of information technology in general and the Internet in particular. We take this as evidence that U.S. inventors seem to be disproportionately relying on knowledge generated within the United States even as the fraction of patents originating overseas has grown. However, this finding could also at least partly be an artifact of patent data. For example, in the absence of inventor/examiner citation distinction for our full sample period, we cannot rule out the possibility that the United States is becoming more specialized in a way not captured by the formal technological classification system. If this were true, we might still observe U.S. patents as increasingly citing other domestic patents in a way not fully reflecting true knowledge diffusion trends associated with national borders. Although we also cannot rule out a possibility that part of the uptick might be caused by some idiosyncrasy in how USPTO functioning has evolved, just the fact that this effect has not declined over time seems remarkable.

Turning to the counterintuitive finding of an independent state border effect, it is worth noting that a few studies (e.g., Holmes 1998) have found state-level effects in other contexts. Belenzon and Schankerman (2012) even document state-level localization specifically for knowledge diffusion, though they examine only patents assigned to universities and hence argue that policies promoting within-state knowledge diffusion from state-funded public universities might be a driver of their finding. Our study reveals that the state border effect is more general, applying even to knowledge arising in private companies. Further insight into the puzzling state border effect will require a comprehensive exploration of considerations such as government support for research, spending on higher education, and policies affecting interfirm mobility of personnel (Marx et al. 2009). Another fruitful research direction might be exploring diffusion of knowledge through localized networks of individuals (Singh 2005, Singh and Agrawal 2011), with a focus on differences in the nature of formal and informal networks operating at different geographic levels and how various networks might have evolved over time in a way that can explain the state-level localization effect (but not the country-level effect) of declining over time. Linking the existence and evolution of such networks to underlying institutional differences across regions would then be a natural next step.

Further exploration of mechanisms like institutional factors and policies seems promising for future research, but we cannot rule out that at least some of the effects we find might not be robust to using alternate research designs. At a minimum, however, our study is an initial inquiry into border-related diffusion effects for flow of ideas, paralleling analogous studies that disentangle border effects and spatial proximity effects for flow of goods in cross-regional trade (McCallum 1995; Wolf 2000; Anderson and Wincoop 2003; Hillberry and Hummels 2003, 2008). Further progress toward unpacking the geography of knowledge spillovers would also help refine existing theoretical models of innovation, entrepreneurship, and growth, ultimately leading to more effective innovation-related policies.

Acknowledgments
The authors thank INSEAD and the MIT Sloan School of Management for funding this research. They are grateful to Ajay Agrawal, Paul Almeida, James Costantini, Iain Cockburn, Pushan Dutt, Lee Fleming, Jeff Furman, Josh Lerner, Ilian Mihov, Peter Thompson, Brian Silverman, and Olav Sorenson for feedback. The authors also thank seminar audiences at Boston University, INSEAD, London Business School, Singapore Management University, and the University of California (Berkeley), as well as participants at the Academy of Management Meetings (2010 and 2012), National University of Singapore Conference on Research in Innovation and Entrepreneurship (2010), Asia-Pacific Innovation Conference (2011), National Bureau of Economic Research Productivity Lunch (2012), and the Georgia Tech Roundtable for Engineering Entrepreneurship Research (2012) for comments. Any errors remain the authors' own.

Appendix. Details of Sample Construction and Weights Calculation

A.1. Basic Choice-Based Sampling
Choice-based sampling involves drawing a fraction (γ) of the "ones" and a smaller fraction (α) of "zeroes" from the population. The probability of a citation *conditional on a dyad being in the sample* follows from Bayes' rule:

$$\Lambda_i' = \frac{\gamma\Lambda_i}{\gamma\Lambda_i + \alpha(1-\Lambda_i)} = \frac{\gamma}{\gamma + \alpha e^{-\beta X_i}} = \frac{1}{1+e^{-(\ln(\gamma/\alpha)+\beta X_i)}}.$$

So the usual logistic estimation would lead to biased results (Greene 2003). Because the functional form is still logistic, one way to correct the logit estimates is to subtract $\ln(\gamma/\alpha)$ from the constant term. However, noting that such a correction is overly sensitive to the assumption of the logistic functional form being completely accurate, Manski and Lerman (1977) suggest instead the WESML estimator obtained by maximizing the following weighted "pseudolikelihood" function:

$$\begin{aligned}\ln L_w &= \frac{1}{\gamma}\sum_{\{y_i=1\}} \ln(\Lambda_i) + \frac{1}{\alpha}\sum_{\{y_i=0\}} \ln(1-\Lambda_i)\\ &= -\sum_{i=1}^{n} w_i \ln(1+e^{(1-2y_i)x_i\beta}),\end{aligned}$$

where $w_i = (1/\gamma)y_i + (1/\alpha)(1-y_i)$. As Amemiya (1985, §9.5.2) demonstrates, consistency of WESML comes from the expected value of the weighted log likelihood turning out to be the same (except for a scaling factor) as the expected log likelihood for the same sample resulting through random (exogenous) sampling. WESML can be implemented using a logistic approach by "simulating" an exogenous sample by weighting each observation by the number of elements it represents from the population (i.e., by the reciprocal of the ex ante probability of inclusion of an observation in the sample). An appropriate estimator of the asymptotic covariance matrix is White's robust "sandwich" estimator. Strictly speaking, WESML is not statistically "efficient" (Imbens and Lancaster 1996). Nevertheless, efficiency issue can be mitigated by employing sufficiently large samples.

A.2. Combining Choice-Based Sampling with Stratification on Explanatory Variables
In basic choice-based sampling, the "zeroes" are all drawn from the $y=0$ population with a uniform sampling rate (α). This approach can be generalized to obtain additional benefits from stratification on key explanatory variables—that is, allowing "α" to vary across different $y=0$ subpopulations (Manski and McFadden 1981; Amemiya 1985, Chap. 9). Let us define z as a label for different strata that takes values $1, 2, \ldots, T$, and note that

$$\begin{aligned}&\Pr(z=z_i \text{ and } y=y_j \mid x=x_i)\\ &\quad= \Pr(z=z_i \mid x=x_i)\Pr(y=y_j \mid z=z_i \text{ and } x=x_i)\\ &\quad= \Pr(z=z_i \mid x=x_i)\Pr(y=y_j \mid x=x_i).\end{aligned}$$

The second equality comes by assuming that the vector $\mathbf{x}$ includes all information about z that affects outcome y—that is, $\mathbf{x}$ is a sufficient statistic for z. (In our settings, this means our controls sufficiently capture technology- and year-related effects on citation likelihood.) Defining the logistic outcome as $v = (z = z_i \text{ and } y = y_i)$ rather than just y, the log-likelihood function with exogenous (random) sample would be

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln[\Pr(z = z_i \text{ and } y = y_i \mid x_i)] \\ &= \sum_{i=1}^{n} \{y_i \ln[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)] \\ &\qquad + (1-y_i)\ln[\Pr(z = z_i \mid x_i)(1-\Lambda(x_i\beta))]\}.\end{aligned}$$

This forms the basis for deriving the pseudolikelihood function for choice-based sampling with stratification. As per the WESML method, each log-likelihood function term needs to be weighted by the inverse of the ex ante probability of that observation being included in the sample. These weights can still be computed as long as the sample as well as the population counts for each stratum are known. Once we have the weights w_{tj} corresponding to $z = t$ $(t = 1, 2, \ldots, T)$ and $y = j$ $(j = 0, 1)$, the required pseudolikelihood function is given by

$$\begin{aligned}\ln L_w &= \sum_{i=1}^{n} \{y_i w_{z_i 1} \ln[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)] \\ &\qquad + (1-y_i) w_{z_i 0} \ln[\Pr(z = z_i \mid x_i)(1-\Lambda(x_i\beta))]\} \\ &= C - \sum_{i=1}^{n} w_i \ln(1 + e^{(1-2y_i)X\beta}),\end{aligned}$$

where $w_i = y_i w_{z_i 1} + (1 - y_i) w_{z_i 0}$ and $C = \sum_{i=1}^{n} w_i \cdot \ln[\Pr(z = z_i \mid x_i)]$.

Since C is independent of β, it can be ignored. Thus, a weighted logistic estimation can again be used, with the weights given by w_i. (Note that the weights now depend not just on y but also on the stratum z_i.)

A.3. Applying WESML to (Extended) Matched Samples

This approach can be extended to matched samples such as the one we have constructed following Jaffe et al. (1993). For a given cited patent, because the matched patent is drawn randomly from the year and technology class of an actual citing patent, we can interpret each {citing year, citing class} combination as a different stratum and calculate the implied sampling rates based on the sample and population counts for each stratum to determine appropriate weights.

However, the matched sample is not representative of the population because the {citing year, citing class} combinations for which no actual citations ("ones") exist are ignored from the point of view of the potential citations ("zeroes"). To ensure the strata are mutually exclusive and exhaustive while still keeping their number manageable, we create (for each cited patent) a new observation by randomly selecting one potentially citing patent for each year (in the 10-year window) belonging to one of the technology classes from which no citation occurs (in that year). The weight for each of these is computed using the implied sampling rates for random draws from these subpopulations.

An example should clarify the sample construction. One of our cited patents is 4,205,881, applied for in 1980 and in tech class 299. It receives two citations during 10 years: from 4,441,761 {year 1982, class 299} and 953,915 {1989, 299}. Therefore, patent pairs (4,205,881, 4,441,761) and (4,205,881, 4,953,915) represent actual citations ("ones") included with a weight of 1 (as we include all citations, i.e., set $\gamma = 1$). In JTH-based matching, citing patent 4,441,761 was matched to control patent 4,402,550 {year 1982, class 299}. In year 1982 and class 299, there were 92 potentially matching patents from which patent 4,402,550 was chosen through a random draw. So the observation (4,205,881, 4,402,550) was included as a control pair ("zero") with a weight of 92. Similarly, citing patent 4,953,915 mentioned above was matched to control patent 4,974,907 {1989, 299}. In year 1989 and class 299, there were 59 potential matches from which 4,974,907 was chosen. So the observation (4,205,881, 4,974,907) was included as a control pair ("zero") with a weight of 59. Finally, for each year 1981 through 1990, we selected a random potentially citing patent, constrained not to be from technology class 299 for the years 1982 and 1989 (as class 299 is already included in finer strata above just for these two years). The range of weights for these 10 observations ended up being between 61,578 and 99,371, depending on the number of eligible patents in the given citing year.

References

Agarwal R, Audretsch D, Sarkar MB (2007) The process of creative construction: Knowledge spillovers, entrepreneurship, and economic growth. *Strategic Entrepreneurship J.* 1(3–4):263–286.

Alcacer J, Gittelman M (2006) Patent citations as a measure of knowledge flows: The influence of examiner citations. *Rev. Econom. Statist.* 88(4):774–779.

Almeida P, Kogut B (1999) Localization of knowledge and the mobility of engineers in regional networks. *Management Sci.* 45(7):905–917.

Amemiya T (1985) *Advanced Econometrics* (Harvard University Press, Cambridge, MA).

Anderson JE, Wincoop EV (2003) Gravity with gravitas: A solution to the border puzzle. *Amer. Econom. Rev.* 93(1):170–192.

Arzaghi M, Henderson JV (2008) Networking off Madison Avenue. *Rev. Econom. Stud.* 75(4):1011–1038.

Audretsch D, Feldman M (1996) R&D spillovers and the geography of innovation and production. *Amer. Econom. Rev.* 86(3):630–640.

Belenzon S, Schankerman M (2013) Spreading the word: Geography, policy and university knowledge diffusion. *Rev. Econom. Statist.* Forthcoming.

Branstetter LG (2001) Are knowledge spillovers international or intranational in scope? *J. Internat. Econom.* 53(1):53–79.

Breschi S, Lissoni F (2001) Knowledge spillovers and local innovation systems: A critical survey. *Indust. Corporate Change* 10(4):975–1005.

Coe DT, Helpman E, Hoffmaister AW (2009) International R&D spillovers and institutions. *Eur. Econom. Rev.* 53(7):723–741.

Duguet E, MacGarvie M (2005) How well do patent citations measure knowledge spillovers? Evidence from French innovation surveys. *Econom. Innovation New Tech.* 14(5):375–393.

Ellison G, Glaeser E, Kerr W (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *Amer. Econom. Rev.* 100(3):1195–1213.

Glaeser E, Kallal H, Scheinkman J, Schleifer A (1992) Growth of cities. *J. Political Econom.* 100(6):1126–1152.

Gompers P, Lerner J, Scharfstein D (2005) Entrepreneurial spawning: Public corporations and the genesis of new ventures, 1986 to 1999. *J. Finance* 60(2):517–614.

Greene WH (2003) *Econometric Analysis*, 5th ed. (Prentice Hall, Upper Saddle River, NJ).

Greene WH (2009) Testing hypotheses about interaction terms in nonlinear models. *Econom. Lett.* 107(2):291–296.

Grossman G, Helpman E (1991) *Innovation and Growth in the World Economy* (MIT Press, Cambridge, MA).

Henderson R, Jaffe A, Trajtenberg M (2005) Patent citations and the geography of knowledge spillovers: A reassessment: Comment. *Amer. Econom. Rev.* 95(1):461–464.

Hillberry R, Hummels D (2003) Intranational home bias: Some explanations. *Rev. Econom. Statist.* 85(4):1089–1092.

Hillberry R, Hummels D (2008) Trade responses to geographic frictions: A decomposition using micro-data. *Eur. Econom. Rev.* 52(3):527–550.

Holmes TJ (1998) The effect of state policies on the location of manufacturing: Evidence from state borders. *J. Political Econom.* 106(4):667–705.

Imbens GW, Lancaster T (1996) Efficient estimation and stratified sampling. *J. Econometrics* 74(2):289–318.

Jacobs J (1969) *The Economy of Cities* (Random House, New York).

Jaffe AB (1989) Real effects of academic research. *Amer. Econom. Rev.* 79(5):957–970.

Jaffe AB, Trajtenberg M (2002) *Patents, Citations and Innovations: A Window on the Knowledge Economy* (MIT Press, Cambridge, MA).

Jaffe AB, Trajtenberg M, Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Quart. J. Econom.* 108(3):577–598.

Keller W (2002) Geographic localization of international technology diffusion. *Amer. Econom. Rev.* 92(1):120–142.

Kerr W, Kominers SD (2010) Agglomerative forces and cluster shapes. HBS Working Paper 11-061, Harvard Business School, Boston.

Klepper S, Sleeper S (2005) Entry by spin-offs. *Management Sci.* 51(8):1291–1306.

Krugman P (1991) *Geography and Trade* (Leuven University Press, Leuven, Belgium).

Lai R, D'Amour A, Fleming L (2009) The careers and co-authorship networks of U.S. patent holders since 1975. Working paper, Harvard Institute for Quantitative Social Science, Harvard Business School, Boston.

Lampe R (2012) Strategic citation. *Rev. Econom. Statist.* 94(1): 320–333.

Manski CF, Lerman SR (1977) The estimation of choice probabilities from choice based samples. *Econometrica* 45(8):1977–1988.

Manski CF, MacFadden D (1981) Alternative estimators and sample designs for discrete choice analysis. Manski C, McFadden D, eds. *Structural Analysis of Discrete Data with Econometric Applications* (MIT Press, Cambridge, MA), 1–50.

Marshall A (1920) *Principles of Economics* (Macmillan, London).

Marx M, Strumsky D, Fleming L (2009) Mobility, skills, and the Michigan non-compete experiment. *Management Sci.* 55(6): 875–889.

McCallum J (1995) National borders matter: Canada-U.S. regional trade patterns. *Amer. Econom. Rev.* 85(3):615–623.

Peri G (2005) Determinants of knowledge flows and their effect on innovation. *Rev. Econom. Statist.* 87(2):308–322.

Romer PM (1990) Endogenous technological change. *J. Political Econom.* 98(5, Part 2):S71–S102.

Rosenthal S, Strange W (2001) The determinants of agglomeration. *J. Urban Econom.* 50(2):191–229.

Rosenthal S, Strange W (2003) Geography, industrial organization, and agglomeration. *Rev. Econom. Statist.* 85(2):377–393.

Rysman M, Simcoe T (2008) Patents and the performance of voluntary standard-setting organizations. *Management Sci.* 54(11):1920–1934.

Saxenian AL (1994) *Regional Advantage: Culture and Competition in Silicon Valley and Route 128* (Harvard University Press, Cambridge, MA).

Singh J (2005) Collaborative networks as determinants of knowledge diffusion patterns. *Management Sci.* 51(5):756–770.

Singh J (2007) Asymmetry of knowledge spillovers between MNCs and host country firms. *J. Internat. Bus. Stud.* 38(5):764–786.

Singh J, Agrawal A (2011) Recruiting for ideas: How firms exploit the prior inventions of new hires. *Management Sci.* 57(1): 129–150.

Sorenson O, Fleming L (2004) Science and the diffusion of knowledge. *Res. Policy* 33(10):1615–1634.

Thompson P (2006) Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *Rev. Econom. Statist.* 88(2):383–389.

Thompson P, Fox-Kean M (2005) Patent citations and the geography of knowledge spillovers: A reassessment. *Amer. Econom. Rev.* 95(1):450–460.

Wolf HC (2000) Intra-national home bias in trade. *Rev. Econom. Statist.* 82(4):555–563.