The Geographic Imagination of Civil War-Era American Fiction

*Matthew Wilkens**

Space is important in literary studies. This was true even before postmodernism's spatial turn a generation ago, and our collective interest in spatial issues has only grown in recent years. Of course, what we mean by space varies widely across the discipline. We have studies of the relationship between literature and geography at scales ranging from the local to the global. We're also interested in the smaller scales of built space and the lived environment. And then there's the longstanding problem of mapping between space and time as organizing principles of narrative and other forms of cultural production.[1]

This variety doesn't imply that we've made a hash of things. On the contrary, I think we've done well, considering the scope of the problem. But in nearly every case, we work in a way that makes some questions much easier to ask and to answer than others. By this I mean that our need to work with individual texts (and other cultural objects) has led us to study first and foremost specific representations of space and geography, which we've then used as symptomatic indices of larger social configurations. So we've become very good at, for example, determining what Kate Chopin's stories reveal about the nexus of sexuality and Cajun regional culture or, in a different register, how the Las Vegas strip reveals the symbolic function of commercial built space. At the same time, it has often been difficult to assess the extent to which these symptomatic readings apply to larger groups of texts and objects, simply because we've lacked the capacity to apply our methods sequentially across, say, all of the fiction written in the nineteenth century.

**Matthew Wilkens* teaches American literature and digital humanities at the University of Notre Dame. His work has appeared in *Contemporary Literature*, *New Literary History*, *Post45*, and the *Los Angeles Review of Books*. He is currently working on a book about geographic space in American fiction.

American Literary History, vol. 25, no. 4, pp. 803–840
doi:10.1093/alh/ajt045

It should be clear that for the most part this is an indictment neither of any individual study nor of literary-critical methods in general, but instead an identification of a horizon of scholarly possibility to which our attention is rarely drawn. We tailor our work to the kinds of questions we can answer straightforwardly and well, which as a practical matter has meant those that yield to close readings of a few texts. This means in turn that potentially interesting questions with which we don't have the ability to grapple haven't been merely ignored or set aside, but have remained largely invisible to us. How contemporary authors—all of them, as a group—make use of interior space, for example, is an issue so far outside our realm of practical engagement as to appear, I suspect, less farfetched than simply absurd.

But it doesn't have to be this way, at least not in the case of geographic space and the limitations imposed by close reading. We now have methods by which to work with large bodies of text and to extract at least some types of spatial information from them. These methods, which involve computational data mining of hundreds or thousands of books, make it possible for us to address large-scale spatial questions, questions of the type that once seemed unthinkable, in new and robust ways. This is especially true because in many cases we can then combine the evidence produced through these new approaches with our well-established critical judgments.

What follows is an example of such hybridized, computationally assisted scholarship. It begins with a question: how can we define and assess the geographic imagination of American fiction around the Civil War, and how did the geographic investments of American literature change across that sociopolitical event? It is, at first order, an intervention in existing debates about space, regionalism, and the dynamics of large-scale cultural change. To preview quickly the most important direct results, we find that there is significant national and international dispersion of geographic reference in American novels written between 1851 and 1875; that the distribution of place references within the United States tracks closely but not perfectly with population; that changes in literary investment in specific places and regions tend to lag changes in population; and that although there are important shifts in the geographic distribution of literary interest occasioned by the Civil War, such shifts are smaller than established theories would lead us to expect.

Beyond presenting this new, broad-based information about the distribution of literary–geographic attention in the mid nineteenth century, my specific claims are three: first, that the New England-centered understanding of American literature and culture that grows out of foundational studies such as F. O. Matthiessen's *American Renaissance: Art and Expression in the Age of Emerson*

and Whitman (1941) and Sacvan Bercovitch's *The Puritan Origins of the American Self* (1975) is misleading and ultimately unsustainable when applied to the full literary field of the mid century. Second, that literary regionalism, as measured by any large uptick in the actual use of regionally dispersed locations, does not arise in the decade following the Civil War, though certain of its roots can be traced at least as far back as 1850. In both cases, our view of the era's literary production should be revised to reflect its significant investment in urban and international locations, as well as the wide range of those locales. Third, that we must rethink significantly our theory of periodizing events in light of the striking continuity of literary–geographic usage across the outbreak and conclusion of the Civil War. This last point isn't an attack on periodization as such, but we do need, I think, to pay much closer attention to the substantive continuities that underlie our narratives of historical evolution, even when those narratives are built around such seemingly obvious breaks as the event of national insurrection.

1. Renaissance and Regionalism

To the very modest extent that we have a consensus view of the way mid-nineteenth-century American fiction as a whole engages geographic space, that view is dominated by two long-established critical narratives: the American Renaissance of the early 1850s and the rise of regionalism after the Civil War.[2] Despite serious reservations concerning aspects of both these narratives, they remain the only widely used large-scale interpretive frameworks through which to understand (even in opposition) the development of American fiction in the period at hand. These critical narratives, in turn, entail relatively straightforward geographic consequences. The American Renaissance, taking place mainly in the Northeast and deeply concerned with issues of American identity and self-discovery, would be expected to have produced texts set in and around New England and, when venturing outside that region, to remain largely within the confines of the US. Exceptions spring to mind, of course; *Moby-Dick* (1851) is set largely outside the territory of the US, for example. But we're describing a situation of overall emphasis rather than of absolute uniformity, and it is difficult to imagine an area that could be said reasonably to surpass New England as the locus of the Renaissance writers' geographic attention.[3]

The roots of this view concerning New England's centrality at mid century lie—beyond Matthiessen's classic *American Renaissance*—in the influential Puritan-oriented theories of American identity advanced by Perry Miller and Sacvan Bercovitch. Miller's *The*

New England Mind: The Seventeenth Century (1939) was a source for Matthiessen, whose study addressed in detail just five writers (Ralph Waldo Emerson, Henry David Thoreau, Nathaniel Hawthorne, Herman Melville, and Walt Whitman), all of whom spent their working lives in or near New England proper and were closely associated with the region's history. It's not the case, of course, that Matthiessen was unaware of the geographic dispersion of American writing by the 1850s; he noted that Hawthorne was nearly the lone proper New Englander among the nation's most distinguished fiction writers and his book touches very briefly on an array of texts beyond those of his primary canon.[4] But Matthiessen's account of the emergence of a distinctively American literature around 1850 was based almost entirely on authors he identified in relation to the nation's colonial origins in the Northeast. Bercovitch went a step further in his *Puritan Origins of the American Self*, arguing that Emerson (in combination with the same set of writers identified by Matthiessen) was "the crucial figure in the continuity of the culture" across the nation's first two centuries and that his essays represented the "fullest expression" of "the American future which [Cotton] Mather proclaimed" (163, 184). This New England genealogy of American literature was, moreover, on Bercovitch's view, absolutely distinct from that of the US South (and of any other new-world colony), where settlers "conceived of their venture as Europeans" (137) rather than as Americans, and which was thus legitimately ignored in any cultural derivation of American national literature as such.

After the Civil War (the narratives continue)—with the basic question of the nation's continuing existence settled—literary regionalism emerged as a way of exploring geographic and cultural difference within a larger political unity.[5] We would thus expect the locations of American fiction after 1865 to show significantly more geographic diversity than those of the antebellum period, including shifts to the south and west, accompanied by the emergence of distinctive local and regional clusters of attention. The growth of regionalism was reinforced by two linked factors, one historical, the other critical. Historically, regionalist writing found a ready audience via the rapidly expanding page counts of postbellum periodicals and their readers' demand for the type of interesting novelty regional and local color writing could provide. In the critical realm, the longstanding association of regionalist writing with the post-Romantic realism of the later nineteenth century further strengthened any narrative that would see regionalism as a phenomenon of the last third of the nineteenth century.[6]

The expected literary–geographic structure over the course of the mid century, then, is one of largely domestic expansion and differentiation. If this is what we find in a survey of the places used in

the period's literature, we'll have another reason to believe that the underlying literary–historical narrative is correct; if not (and indeed, it is not), we'll have a new piece of evidence in light of which to consider changes to that narrative.

2. The Raw Numbers

The results presented here are based on an analysis of more than 1,000 novels by American authors published in the US between 1851 and 1875, from which every explicit mention of a geographic location has been programmatically extracted and mapped. Information about the texts included and methods employed can be found in the technical appendix at the end of this article. The appendix also includes discussions of the unique cultural position occupied by novels during the period, of the quality and limits of the data involved, and of the challenges unique to corpus-based analysis.

It is possible, if eventually unwieldy, to list the locations identified in the corpus by the administrative division to which they belong. Tables 1 and 2 contain the most frequently occurring nations, US states, and cities over the full range of publication dates, 1851–75. Note that in every case, lower-order geographic areas have been included in the count of higher-order areas. So, for instance, the count for the US includes references to the nation as a whole, but also to all cities and states within the US; likewise, the count for New York City includes specific neighborhoods and landmarks within the (modern) city limits.

Table 1. Most frequently occurring nations and US states, 1851–75

Nation	Count	State	Count
US	83,412	New York	14,980
Great Britain	13,296	*None/multiple*	14,121
France	7,730	Virginia	8,256
None/multiple	7,523	Massachusetts	6,734
Italy	5,920	Washington, DC	4,652
Israel/Palestine	2,628	Pennsylvania	3,829
Canada	2,040	California	3,205
Germany	1,874	Louisiana	2,394
India	1,645	Ohio	1,972
Ireland	1,424	Texas	1,826

Counts include lower-order locations.

Table 2. Most frequently occurring cities, 1851–75

City	Count	City	Count
New York	10,240	Baltimore	706
Washington, DC	4,171	Niagara Falls, NY	694
Boston	3,995	San Francisco	682
Paris	3,431	Fairfax, VA	630
London	3,332	Saratoga Springs, NY	625
Rome	2,179	Newport, RI	614
Philadelphia	2,046	*Naples*	559
New Orleans	1,579	Albany, NY	472
Richmond	1,164	*Venice*	442
Jerusalem	929	*Havana*	419
Charleston, SC	876	St. Louis	405

Italics indicate cities outside the US. Counts include lower-order locations (for example, neighborhoods and landmarks).

Given the geographic nature of the data, maps offer a more comprehensive—if necessarily less precise—overview of the results. Figures 1 and 2 collect city-level data across the corpus for locations worldwide and within the US, respectively. As Figure 1 and Table 1 make clear, US locations constitute a clear majority of all named-place occurrences in the corpus (58% of overall occurrences and 61% of occurrences isolatable to a single nation). But a significant fraction of the named locations in the corpus also fall outside the US, a fact considered at length in the following section. Figure 2 shows the distribution of US city-level locations, revealing a preponderance of literary–geographic occurrences in what we would now call the Northeast corridor between Washington, DC, and Boston, but also sizable numbers throughout the South, Midwest, Texas, and California.

Turning to regional and national data, a handful of frequently occurring locations and areas tend to dominate each category. For this reason, unscaled choropleth maps (that is, maps shaded according to the value of an underlying variable such as occurrence count) reveal very little information, showing, for instance, only a sufficiently high concentration of US locations as to swamp almost all visible variation in other nations. Log-scaled counts, as shown in Figures 3 and 4, are a more useful alternative, provided we remain aware of the range compression they introduce.

The distributions illustrated in Figures 3 and 4 are interesting in their own right and are discussed below, but they mask several important underlying phenomena. First, they do not reflect any

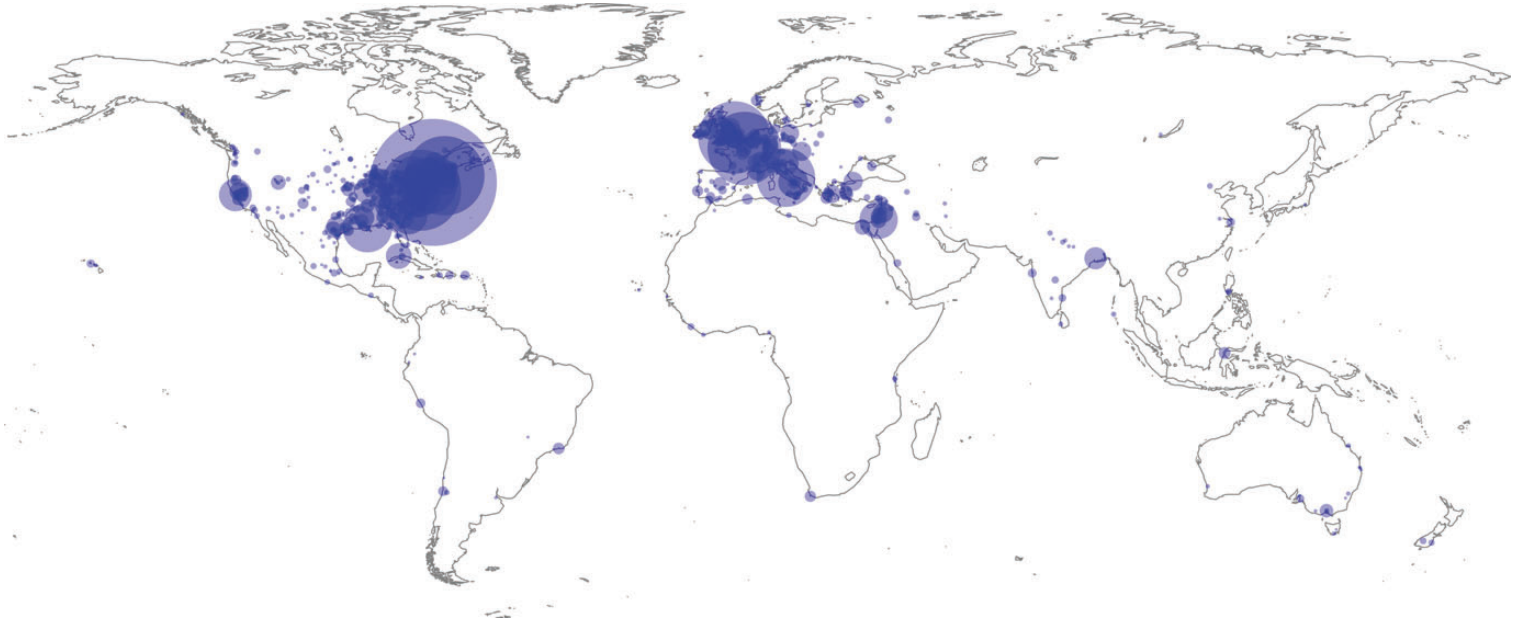

Fig. 1. Worldwide city-level locations in the full corpus. Marker areas correspond to fraction of total occurrences.

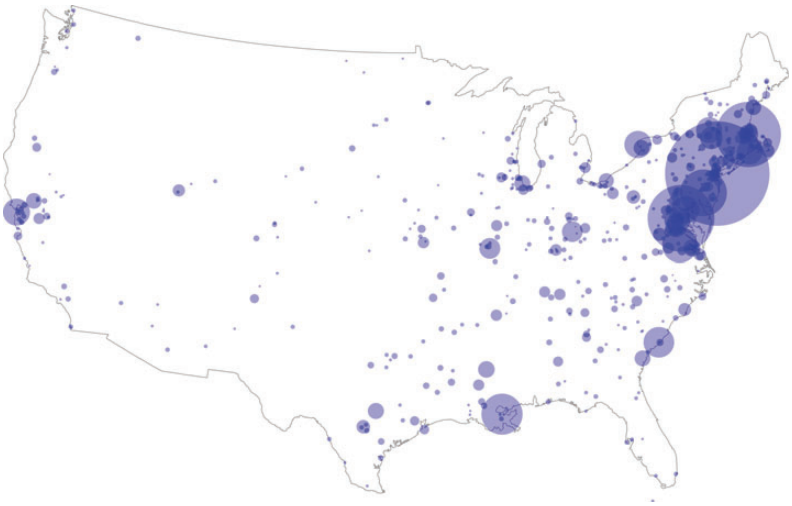

Fig. 2. US city-level locations in the full corpus. Marker areas correspond to fraction of total occurrences.

temporal changes in geographic distribution that may have taken place during the generation or so represented in the corpus, especially across the event of the Civil War; second, they don't account for the changing and very unequal populations of the nations and US states in question. It is possible, of course, to disaggregate raw counts by year and to track their changes across the period; Figure 5 illustrates such an approach, showing the movement of the center of gravity of both human population and literary locations within the US before and after the Civil War.[7] But raw counts alone provide results that are generally difficult to interpret, both because the counts for any one year can vary widely (hence the aggregation into pre- and post-1861 periods in Figure 5; this is especially true in the case of smaller and less frequently invoked regions) and because nearly all regions show an upward trend over time as population and publication volume increase. In fact, population and location occurrence counts are strongly correlated at the state level ($R^2 = 0.59$, meaning that population accounts for about three-fifths of the observed variation in occurrence counts) as shown in Figure 6.

A more useful approach is to compare the fraction of all named-location occurrences that fall within a region to that region's fraction of an overall population. In the case of the US, reliable historical population data at the national and state levels are readily available via the census.[8] In order to compare the number of literary occurrences with widely varying state populations, however, we require (in addition to aggregation or other smoothing of volatile data) an approach that strikes a reasonable balance between large percentage shifts in small numbers and large absolute shifts in large

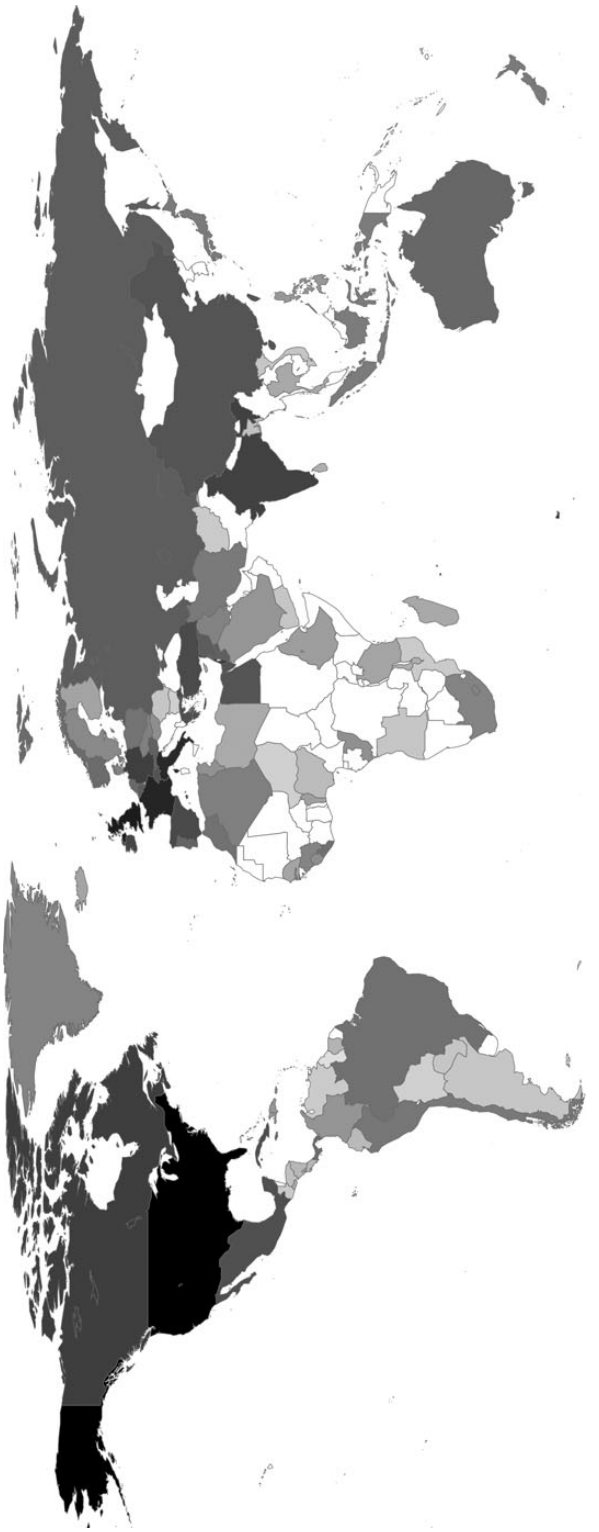

Fig. 3. Log-scaled counts of named locations by nation, 1851–75. Darker shading represents regions containing more occurrences of named places.

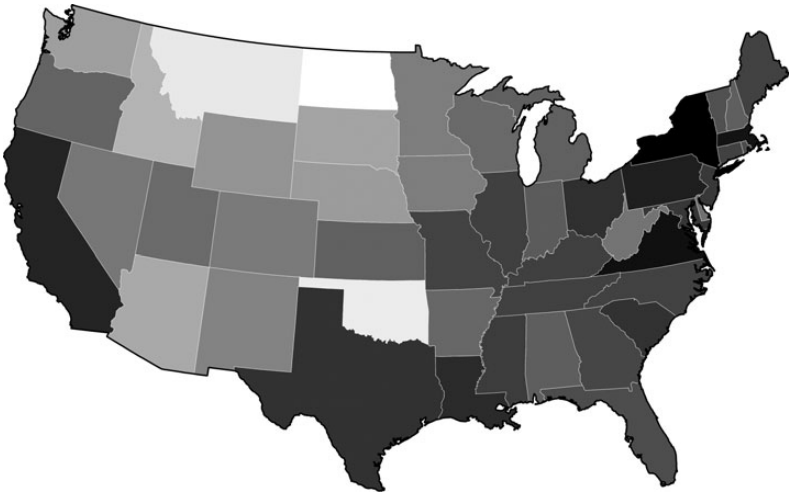

Fig. 4. Log-scaled counts of named locations by US state, 1851–75.

Fig. 5. Geographic mean of US population (squares) and of literary named places in the US (circles), showing westerly trend over time. Note that census dates do not correspond exactly to ranges of literary publication dates.

numbers.[9] One such method is Dunning's log-likelihood, a statistical test based on the likelihood that the difference between two measurements is the result of random fluctuation.[10] A comparison between mean state and territory populations in the period 1850–80 and total

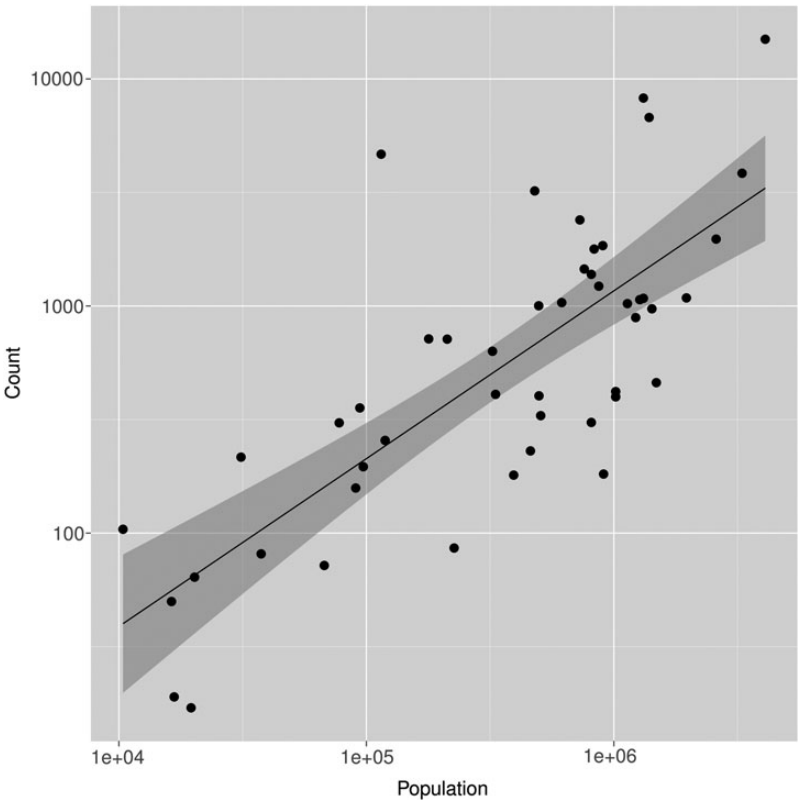

Fig. 6. Log-scaled counts of state-level named locations as a function of log-scaled average state population, 1850–80, with linear regression (performed on log-scaled values; $R^2 = 0.59$*).*

counts of named location occurrences in the corpus using Dunning's method is shown in Figure 7 and summarized in Table 3.[11] Extended discussion of these results follows below; for the moment, note the presence of Ohio among the most statistically underrepresented states despite its relatively high raw occurrence count (see Table 1).

Figure 7—which a colleague once suggested should be titled "The Midwest Has Always Been Boring"—gets us halfway to the goal of a measure that's responsive to differences in population and to changes over time. To introduce temporal resolution, we need to disaggregate location occurrence counts and population figures into at least two periods, and then calculate the differences in log-likelihood across those periods. Figure 8 and Table 4 show the results of such a procedure using two temporal bins, 1851–60 and 1861–75, divided roughly by the outbreak of the Civil War.[12] What the figure and the table reflect are those states that underwent the largest *changes* in their degree of over- or underrepresentation before and after 1861.

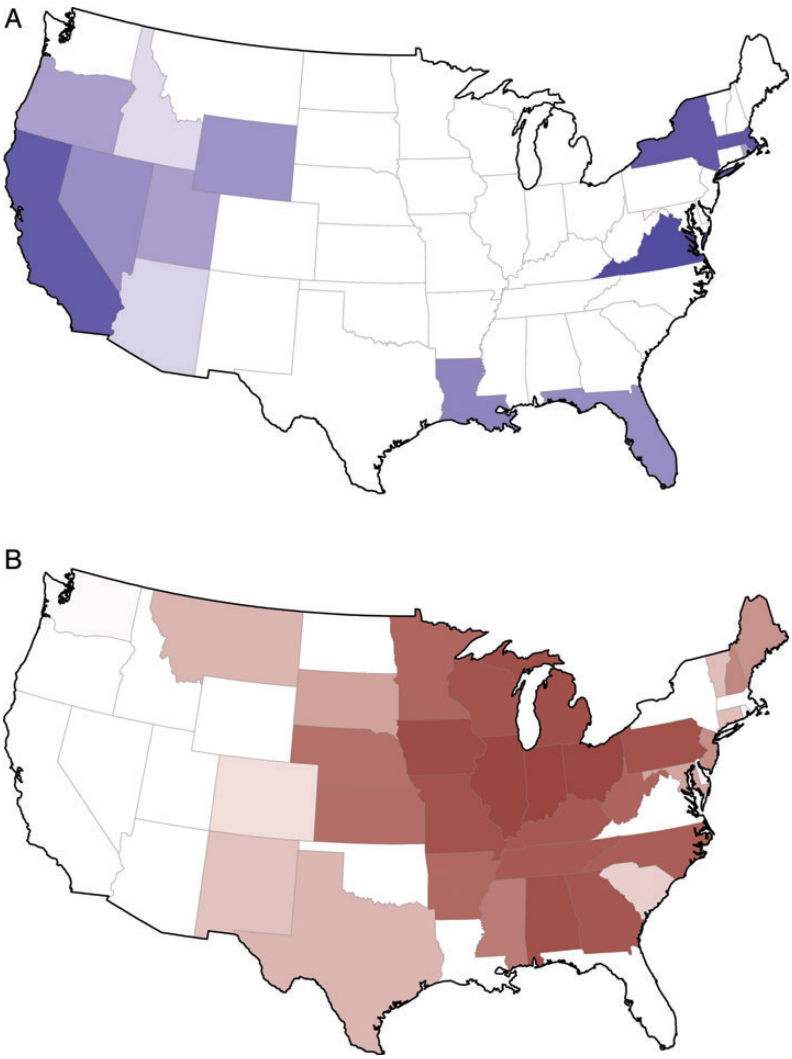

Fig. 7. Dunning log-likelihood values for named-location counts in the full corpus measured against mean state populations, 1850–80. (a) States overrepresented relative to their populations; (b) underrepresented states. Darker shades indicate larger absolute values, hence greater under- or overrepresentation.

These results are discussed at length below. For the moment, note that because log-likelihood values respond to the ratio of two independent measures—population and named-location occurrence count—the changes summarized in Figure 8 and Table 4 represent different kinds of underlying phenomena. On the one hand, states with relatively little population growth will show large changes in Dunning values (that is, in their degree of representation) only if the rates at which locations within those states are used in the literary corpus change significantly. This is particularly true for large,

Table 3. Most over- and underrepresented states in the literary corpus compared to mean state population (1850–80) as calculated by Dunning's log-likelihood method (in thousands)

Overrepresented	Dunning values	Underrepresented	Dunning values
Washington, DC	17.9	Indiana	−3.9
Virginia	6.2	Illinois	−3.7
Massachusetts	3.1	Ohio	−3.5
New York	2.9	Iowa	−2.9
California	2.7	Michigan	−2.4

frequently occurring states like New York and Virginia that are relatively impervious to the vagaries of individual books set within their borders. Fast-growing states, on the other hand, require large increases in their number of literary–geographic occurrences merely to maintain their initial ratios of literary mentions to population size. For this reason, many of the largest decreases in degree of representation over the period are seen in states and territories with modest initial populations and high levels of population growth between 1850 and 1880 (including Minnesota, Oregon, Iowa, and Texas; see Figure 9 for a comparison of state-level populations over the period). Infrequently occurring states are also especially prone to large fluctuations due to their presence or absence in individual texts. All of this is to say that changes in both demographics and literary attention can and do drive changes in the degree of under- or overrepresentation of states as measured by the Dunning method, and that we must consider both factors when assessing the numbers presented in Figure 8 and Table 4.

3. Discussion, or, What the Numbers Mean

The data presented above are valuable at first order as an unprecedented inventory of the geographic locations used in American fiction around the Civil War. For the current project, they also allow us to form a newly robust sense of the geographic imagination of mid-nineteenth-century fiction and to evaluate the related literary–historical narratives (the American Renaissance, the rise of regionalism) discussed in the introduction. To recall, we expected to find relatively sharp differences in geographic usage between fiction written before and after the war, specifically a high concentration of literary locations in the northeastern US before 1865, followed by a marked transition to regional dispersion after that date. In both periods, however, the literary and cultural emphasis was thought to

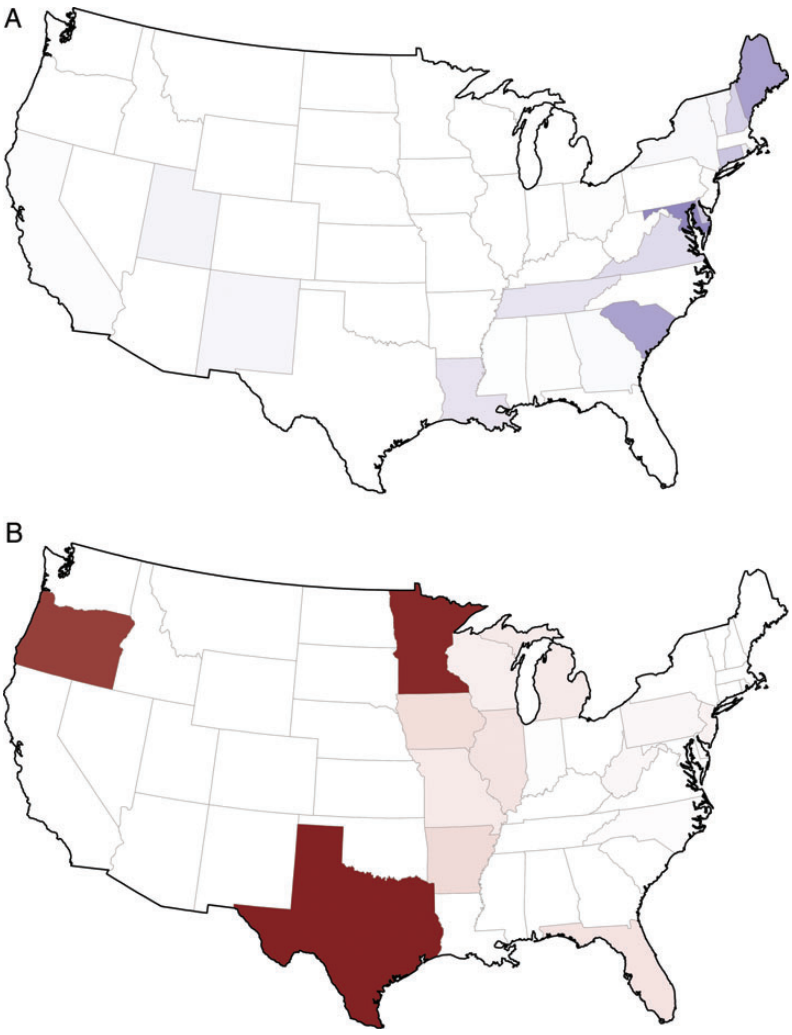

Fig. 8. Changes in log-scaled Dunning log-likelihood values for named-location counts in the periods 1851–60 and 1861–75 against 1850 and 1880 state populations, respectively. (a) States with increased degrees of representation (relative to population) after the Civil War; (b) states with decreased representation. Darker shades indicate larger absolute values, hence larger changes in degree of under- or overrepresentation.

be on American themes, hence also on locations within the US rather than those outside the nation's borders.

The US and the World

Considering first the gross distribution of literary locations at the global scale, two features stand out. As noted above, locations within the US predominate, accounting for just under 60% of the

Table 4. Largest changes in degree of literary representation relative to population across the Civil War as measured by changes in log-scaled Dunning's log-likelihood values for named-location counts by state in the periods 1851–60 and 1861–75 against 1850 and 1880 state populations, respectively (minor discrepancies in Δ are due to rounding for presentation)

Increased representation	Pre-1861	Post-1861	Δ	Decreased representation	Pre-1861	Post-1861	Δ
Maryland	−2.33	0.76	3.08	Texas	2.88	−2.48	−5.36
Maine	−2.47	−0.08	2.38	Minnesota	2.24	−2.96	−5.20
South Carolina	−1.78	0.44	2.35	Oregon	2.96	−1.71	−4.67
Connecticut	−1.59	−0.20	1.40	Arkansas	−1.98	−2.85	−0.87
Delaware	−1.06	0.33	1.39	Iowa	−2.57	−3.34	−0.77

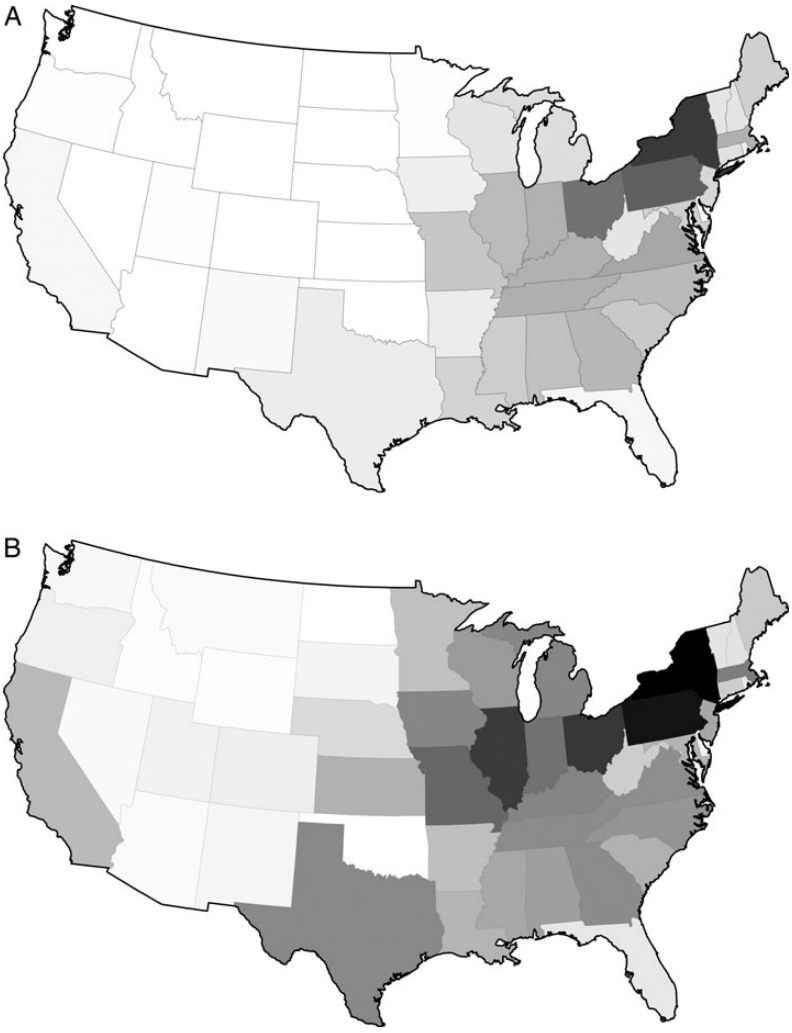

Fig. 9. Population of US states in (a) 1850 and (b) 1880.

named-place occurrences in the corpus (see Figures 1 and 3 and Tables 1 and 2). This is a significant fraction and the extent to which American authors use American locations can be striking; among the hundreds of occurrences of "Cambridge," for instance, no more than a handful refer to the English university or town. Yet international locations are by no means rare, making up over 40% of total occurrences. And while some of the sites of international attention are more or less expected—Britain, France, and Italy are the three most frequently occurring nations outside the US, accounting together for fully one-fifth of all occurrences traceable to any nation—many range far afield. Africa and its sublocations occur 2,458 times; India, 1,652 times; China, 679 times; and Russia, 461 times. This means that African locations collectively fall between London and Rome in

terms of frequency, that India and New Orleans receive similar numbers of uses, and that China occurs about as often as does San Francisco. All told, locations outside the US, UK, France, and Italy make up over 18% of total national- and lower-level occurrences, a sum equal to New York, Washington, Boston, Paris, and London combined.

Also striking is American authors' frequent invocation of locations associated with the Bible and with classical antiquity. Locations in present-day Israel, Palestine, Syria, Egypt, Turkey, and Greece together account for 5,754 occurrences, or about 4% of the total. (Italy is more complicated, being split between contemporary and classical uses.) Nearly two-thirds of all volumes contain at least one use of a location in those countries. It's no surprise, of course, that biblical and classical references were common in mid-nineteenth-century fiction, but the fact that such locations make up roughly 1 out of every 25 named places in the corpus is a remarkable indication of the extent to which American literature at the time operated through Christian and classical frameworks, producing a literary-imaginative geography that was shaped in significant and visible ways by the spaces of the Bible and the classics.[13] This is true even when such locations were used allusively or metaphorically, rather than as the sites of plot-level action; what's important is that they were at the time common elements of the structure of geographic attention, that is, places marked prominently on an American map of the imaginary world.

It is also helpful to have this background in mind as we evaluate the geographic usage of canonical authors, on whose work much of our received wisdom about the period is based. Hawthorne, for instance, though not systematically averse to international locations (see in particular *The Marble Faun* [1860]), uses antique locations much less often than his contemporaries; they account for just 2% of his place mentions and occur in only half of his texts represented in the corpus.[14] Herman Melville, by contrast, leans more heavily on the ancient world, invoking its locations at more than three times Hawthorne's rate.[15] It may be true, as Matthiessen long ago noted—transporting Samuel Taylor Coleridge's Romantic aesthetic to an American context—that Hawthorne's work is distinguished by its reliance on allegory, Melville's on symbolism. But if so, Hawthorne's allegory (a form disparaged for its frequently plodding biblical explicitness) was accomplished with remarkably little help from direct reference to the Holy Land, while Melville's symbolism was accompanied by a surprising number of such unsymbolically direct invocations, from Jaffa, Judea, and Kedron in *Moby-Dick* to Jerusalem, Canaan, and Egypt in *Pierre* (1852).

There is, however, an important difference between the various classes of locations: so-called exotic places, those with which

authors and readers were presumably unfamiliar, tended to be used with much less geographic specificity than their more commonplace counterparts. In the case of the US, for instance, 57% of the location occurrences in the corpus are at the city level or lower. In Africa, the figure is 17% (the continent name itself accounts for 28%; the remaining 55% are country names). In India, cities account for 32% of references; in Russia, 25%; in China, just 11%. The same trend is visible within the US: in Pennsylvania and Massachusetts, two Eastern states with large publishing industries and comparatively advanced literary markets, locations at the city level and below account for 77% and 88%, respectively, of all occurrences. In Iowa and Oregon (the latter still a territory for part of the period in question), the numbers are radically lower: 18% and 13%, respectively. There are exceptions, of course, especially when a single city dominates attention devoted to a larger region, in effect taking over synecdochically the role of the region as a whole (as is true of New Orleans vis-à-vis Louisiana, for instance), but in general we can say that the further geo-imaginatively afield a text ventures, the less spatially specific it becomes.

With this caveat in mind, the degree to which American fiction around the Civil War makes use of widely distributed foreign locations remains unexpectedly large. Elsewhere I have described the period's literary production as diversely outward looking in ways that should unsettle our standard accounts of the period.[16] This is true, I think, not because every mention of a location outside the US (nor within it, for that matter) indicates a deep investment in that place, but because overall patterns of attention matter deeply when we're considering large-scale views of an entire culture's literary output. Whatever it was that American fiction was doing around the Civil War, it was doing it by way of a system of geographical references that fell almost as often outside the borders of the nation as within them.[17] The numbers involved are too large for this distribution to be either accidental or incidental; there are 299 volumes (out of 1,050) that mention at least one location in China, 187 that refer to Russia, 221 to Cuba, 669 to France, 340 to Germany, and 805—more than three out of every four—contain uses of British place names. It would be both feasible and rewarding to undertake a full study of the ways in which any one of these nations was used in the literature of the period, or of their collective use in a single text or author, but to do so would miss the weight of the present point. The remarkably extensive use of foreign locations across most books of fiction published in the period indicates a literary-cultural investment that is at once vitally important and impossible to demonstrate via almost any close reading, real or imagined.

The remarkably extensive use of foreign locations across most books of fiction published in the period indicates a literary-cultural

investment that is at once vitally important and impossible to demonstrate via almost any close reading, real or imagined.

To explain this international diversity is a compelling challenge. Some of it surely derives from simple precedent and historical influence. The American book trade throughout the nineteenth century relied heavily on cheap (and almost universally unauthorized) reprints of British texts, which were both popular and widely circulated.[18] If, as seems reasonable to assume given the results in the American case, British texts featured predominately British and Continental locations, and if those texts served as paradigms for American authors and readers alike, works produced in the US and sold into the same market would be expected to display any number of similarities to their British counterparts, patterns of geographic usage among them. It is also the case, of course, that the US was a far less politically and economically consequential nation during the period in question than it is today, a fact that we should expect to have driven attention toward locations of greater perceived significance.[19] This is to say—notwithstanding Bercovitch, Matthiessen, Miller, and, ultimately, Emerson—that the geography of American fiction was almost certainly often imported from Britain and the Continent. This was true not just in the colonial period, when it might have been expected (though, again, Miller and Bercovitch suggest otherwise), but through the generation or more following Emerson's and Whitman's explicit calls for an American literature that would "speak our own minds" free of "the courtly muses of Europe" (Emerson 69) so that Americans might produce what Whitman called "the great psalm of the republic" (619).

We might also discern issues of specifically American concern or identity in some uses of foreign locations, recognizing that any sense of national definition will necessarily be formed in part through contrast with the imagined traits and experiences of others. Nevertheless, the details of these contrasts matter in two ways. First, the raw frequency with which foreign locations are used suggests that even if some portion of those uses refer finally to domestic issues, those issues were frequently framed in international terms.[20] Second, the world is not undifferentiatedly other in this set of formative American texts, which make consistent use of a huge array of foreign locales; references to London and Paris alone were plainly not sufficient to explore the larger world, nor to assess the many potential American relations to it. This is to say that even in those cases in which foreign locations were used for what we might call domestic ends, the range of such locations is sufficiently large as to suggest meaningful variety in both the nature of those ends and the geographic–imaginary frameworks through which they are achieved. To cite only one set of conceptual examples, consider the difference between the direct biblical invocations of Jerusalem that permeate much of the period's writing (Lizzy Bates's quotation of Psalm 125

—"As the mountains are round about Jerusalem, so the Lord is round about his people"—in *Had You Been in His Place* [1873] is typical), related but figurative uses of the same (Henry Ward Beecher compares a view of Amherst College to that of "Jerusalem afar off" in his widely circulated *Norwood* [1868]), Twain's satirical use of the Holy Land in *Innocents Abroad* (1869), Melville's deployment in *Pierre* of China as the most remote location a voyage might achieve, and the mixture of envy and condemnation of Europe's great cities that comes from describing New York as at once "the Brooklyn . . . of London and Paris" (Newell 257), the new Vanity Fair (Curtis 34), and the "Future" to London's "Past" (Kirkland 154). This is not a list of locations characterized by simple and repetitive alterity. Its variety and the much larger varieties of usage for which it stands renders untenable any claim that American fiction of the period was concerned only with separating American from foreign experience as such.

Domestic Locations over Time

Shifting attention to the locations that fall within the borders of the modern US and to the modest majority of all named location occurrences found there, we note several important features.

In partial agreement with our expectation concerning the imaginative prominence of New England during the early portion of the period in question, we find significant numbers of named location occurrences in that region. New England is prominent in Figures 2 and 4, though the largest numbers of occurrences on the east coast—apart from Boston—fall further south in New York, Pennsylvania, Washington, DC, and Virginia. Indeed, New England locations in sum account for just 15% of all US occurrences over the full period.[21] Still, that fraction is well above the region's population, which accounted for 9% of the average US total between 1850 and 1880.

Changes across time and variations over space complicate the assessment of New England's role. Of the six states that make up the region, only Massachusetts and (to a much smaller extent) Rhode Island are overrepresented compared to their populations. Within Massachusetts, more than three quarters of references are to Boston, Cambridge, or to the state itself; in Rhode Island, the resort town of Newport—among the century's most fashionable summer retreats—is the entire story. It appears, then, that New England's overrepresentation is primarily a function of Boston's outsize role as a location in and of the period's fiction. In this role, Boston is similar to other major cities (of which it was then between the third and seventh largest in the US), which often dominated their states and regions in the literary imagination.

How did New England's prominence shift across the Civil War? The answer isn't entirely straightforward; literary locations in the region accounted for 16.7% of all US occurrences in the decade before 1861, then fell slightly—to 14.1%—from 1861 through 1875. But New England's share of overall US population was dropping more quickly at the time (to 8.0% in 1880 from 11.8% in 1850).[22] So the extent to which New England was overrepresented in American literary output relative to its population share actually increased after the war, despite the fact that the region's places made up a smaller fraction of those used in postbellum American fiction. It's tempting to assume that Boston, the region's primary literary–geographic driver throughout the period and home to nearly all its major publishers, followed the same pattern, but this isn't quite right; unlike the rest of New England, Boston grew at a rate well above the national average (in keeping with the country's increasing urbanization), yet the rate at which the city was referenced in American fiction after 1861 declined to 4.4% of all named locations from 5.4% in the antebellum period.

What do these competing trends imply for our larger investigation? For one thing, it's true that New England was overrepresented relative to its population in the decade we associate with the American Renaissance. But the region was also overrepresented—even more overrepresented, in fact—in the years after the Civil War, when we would have expected it to have fallen sharply in literary prominence. Boston, however, performed more as our conventional literary–historical narratives would have predicted, making up a meaningfully smaller fraction of total literary location uses after the war even as the city itself nearly tripled in size.[23] In either case, however, we should bear in mind both that the absolute magnitude of the changes involved were relatively small (on the order of a percentage point or two) and that New England, despite its overrepresentation, nevertheless accounted for only some 15% of all location occurrences in the corpus. The latter number may or may not ultimately strike us as importantly large, but it is certainly not the case that American fiction around the Civil War was predominantly set in, nor invoked, New England locales.

Similarly complex trends with respect to urbanization, regional dispersal, and overall distribution of location occurrences are visible outside New England. We know, for instance, that cities were growing rapidly throughout the period. In 1850, 15% of the US population lived in urban areas; a generation later, in 1880, the fraction of city dwellers had nearly doubled.[24] And while the biggest cities were driving much of that growth, small cities were expanding even more quickly, so that the portion of all urban residents living in the 20 largest cities fell to 44% in 1880 from 53% in 1850. In an

economically and demographically mature nation like the contemporary US, it can be easy to forget this fact, that urbanization can proceed not just by moving people from farms to existing cities, but also by founding (many) new cities. The process might be described, somewhat paradoxically, as dispersed aggregation: people were moving rapidly into larger groups, but the number of such groups was increasing just as quickly, with significant population centers found scattered throughout areas where none had previously existed. Even as the nation was becoming more urban, there were also many more people in places that had been, until a decade or two earlier, essentially rural.

If these developments had counterparts in the fiction of the period, we would expect to find analogous features in the use of named literary locations. Indeed, meaningful similarities do exist, but we must take into account an important difference between the literary and demographic cases. Where the US population was growing exponentially through the period—and through most of the nineteenth century—the annual volume of published fiction was essentially flat (with the exception of the first years of the Civil War, when it fell sharply; see Figure 10).[25] Since the rate at which named locations were used in the period's fiction also remained roughly constant, this means that there was a finite amount of literary–geographic attention (as measured by the number of place-name references) to go around. If Chicago occurred more frequently, some other city or cities had to appear less often. Thus if the number of cities that fell within the literary ambit was rising after the Civil War (as the regionalist theory would lead us to expect), there had to be cities that were receiving less literary attention, whether their populations were growing or not.

So we want to know, first, whether the number of unique locations was rising after the war and, if so, whether the attention devoted to those new locations was being diverted from the major prewar sites (in which case we would have evidence of growing regionalism) or from existing second-rank locations (in which case we would observe a shift in the specific areas of regional investment, but not growth in its overall extent). Concerning the number of unique locations used after 1860, we do observe a modest increase, on the order of 1.5% above what antebellum usage would lead us to expect.[26] From where, then, was the attention to these new locations drawn? Within the US, the percentage of all locations made up by the dominant cities was lower after 1860 than it was in the preceding years (by about 3 percentage points in absolute terms, or 6% to 8% of the antebellum fraction, depending on whether we consider the 50, 20, or 10 most frequently occurring cities). This means that although the largest cities were growing very rapidly—many of them at least tripled in population between 1850 and 1880—the scarce resource of

literary attention was increasingly directed elsewhere. Although the size of the shift was modest, large cities were, at some level, increasingly out-competed by secondary and tertiary locations in the postbellum literary imagination.

We can see the results of this phenomenon most easily in books that focused on newly developing regions like California, where every city was necessarily a minor one as measured by historical standing. Twain's and Bret Harte's Western writings are typical of this case, as are lesser-known works by A. J. Cline and Archie Argyle that devote extended attention to locations in the Sacramento Valley. In the east, Harriet Beecher Stowe is perhaps surprisingly indicative, demonstrating in her postwar fiction not an abandonment of Boston as a prime regional site, but an increase in the variety and number of other Massachusetts locations used, from Sudbury and Worcester to Dedham and Nahant. Alternatively, explicitly regionalist or protoregionalist texts such as Beecher's *Norwood* (written by Stowe's brother and mentioned above in connection with its telling use of Jerusalem), subtitled "Village Life in New England," focused to a much greater extent on minor towns, including Amherst, Holyoke, and Pittsfield.

At the state level, we have noted already the Midwest's underrepresentation relative to its population (see Figure 7). The Midwest was unique in its uniformity of underrepresentation; it was the only region of the US in which not a single state was overrepresented on average in the corpus. And the degree of literary disinterest in the Midwest grew over time, with only Ohio becoming even slightly less underrepresented relative to its population after the Civil War than it was before 1861. The West and Southwest, on the other hand, were consistently overrepresented relative to their miniscule populations, but only California was strikingly and consistently so. Oregon attracted significant relative interest during the heyday of westward migration via its namesake trail,[27] but dropped into anonymity after the war. Mormonism—hence also Utah—was an object of curiosity (when not moral panic) throughout the period (Wesley Bradshaw's polemical *Brigham Young's Daughter* [1870] is typical, as is the milder curiosity of Artemus Ward and of Twain in *Roughing It* [1872]).

Results were mixed in other regions, with states often diverging widely from their neighbors. On the east coast, it is tempting to attribute New York's consistent overrepresentation, like Massachusetts's, to the presence of a single major city that played an outsize role in the contemporary literary imagination. That may be true, but then Pennsylvania and Virginia are more difficult to explain. Pennsylvania was a large state, the second most populous in the country (behind New York), and had in Philadelphia a large, important city, one that was also home to a major portion of the nation's publishing industry at the time. Yet Pennsylvania was significantly underrepresented relative

to its population both before and after the Civil War. A major city alone appears not to have been sufficient to assure disproportionate literary interest, especially if that city lacked the iconic status of New York and the remainder of the state, then as now, was more closely associated with the Midwest than with the east coast. Virginia, on the other hand, was strongly (and increasingly) overrepresented, yet lacked a single city among the nation's 20 largest.[28] So neither were major cities a necessary condition of literary interest.

What, then, was driving literary–geographic investment at the time? Simple population distribution was plainly a major piece of the answer, but it's also clear that population figures weren't the whole story. While a full account of the underlying factors would eventually devolve into hundreds or thousands of individual readings, there are two issues that merit special attention. First, there often exists some lag between changes in population and corresponding shifts in literary attention. Among the large negative changes in degree of representation shown in Figure 8, for instance, many appear to be the result of such a lag—states that didn't so much sink into obscurity as fail to keep pace in literary usage with their rapidly growing populations. Of the 17 states and territories that experienced declining degrees of literary representation relative to their populations after the Civil War, 13 at least doubled in size and 6 quadrupled or more. Yet only one, Oregon, had a significant decline in its raw number of location occurrences, indicating that literary attention had well and truly turned elsewhere. Among the 20 states that showed rising relative representation, on the other hand, only 3 at least doubled their population.

To illustrate the issue more concretely, consider the cases of New Orleans and Chicago. On the eve of the Civil War, the two cities' populations were roughly equal, a fact that would often predict equivalent degrees of literary usage.[29] But the two cities differed sharply in their histories and were growing at vastly different rates. In 1860, New Orleans was long established and enjoying consistent, linear population growth; Chicago was less than 25 years old and on its way from a nonentity to the nation's second-largest city—with more than a million residents—by 1890. Which location would we expect to have received greater literary attention, the established Southern port or the explosively rising northern city? The correct answer isn't obvious. New Orleans was an established city and an important one, the largest in the South after 1830, economically central throughout the period and a pivotal location during the Civil War (though that fact wouldn't have been reflected directly in books published before 1861, and probably not for some time thereafter). Chicago had none of this history, but it did have the potential advantages of novelty, dynamism, and future significance. If literary

interest were driven in important ways by the search for new experiences and new locations, or by a quest to present or to understand the forces shaping the evolution of the nation, we might expect Chicago to have attracted disproportionate attention at the precipice of its remarkable growth. To this we might add the observation that some literature obviously *is* motivated by such concerns, as in the case of travel and adventure stories (recall the success of Melville's early work, and of Twain's) and as evinced in part by the wide distribution of national and global locations discussed above. Finally, we would add that writing and reading about a place represent significantly lower investments of time and resources than physically moving to that same location, a fact that would be expected to favor literature as a leading indicator of demographic shifts rather than a trailing one.

Despite these competing factors, the record in the corpus is lopsided. New Orleans was vastly more represented than was Chicago in every year before the 1870s. Chicago was never mentioned more than 10 times in any year's literary output until 1867 and didn't surpass New Orleans in any year before 1872. Over the two decades, 1851–70, New Orleans was the more frequently occurring by a factor of more than 15. During the last five years studied, 1871–75, New Orleans continued to lead, though by a smaller margin.[30] In combination with the state-level data discussed above and presented in Figure 8 and in Table 4, this discrepancy suggests that literary attention does not as a rule significantly anticipate shifts in population and in fact generally trails those shifts. This may be due, in some part, to the temporal asymmetry introduced by historical fiction, but the more likely large-scale explanation concerns what we might call the banality of most fiction, especially outside those canonical texts that we continue to study in depth a century and a half later; books are for the most part written about places, events, and topics that have been written about before—about things and places with which readers are more or less familiar. In other words, literature pays attention to the things to which it is already accustomed to paying attention. Such conservatism of attention isn't necessarily bad or unproductive; it is a variety of the same impulse that gives rise to and maintains literary genres. But it does affect the ways in which the literary field as a whole reacts to novelty, acting as a brake on major shifts in content, style, and setting. To this we should add the fact of temporal lags induced by time spent writing, editing, in production, in distribution, and so forth, all of which compound the apparent inertia of literary attention even in cases of comparatively rapid evolution.

Despite the conservative lag in literary responses to demographic shifts, there obviously exist cases in which the locations of literary investment change both rapidly and in conjunction with factors other than population. We've seen a handful of these already:

the end of the wagon-trail era in Oregon, a mounting fascination with Mormonism and its home in Utah, the California gold rush. But most striking among such instances are those related to war and conflict. Consider Texas, which experienced the largest drop in degree of representation relative to population among all states across the Civil War. True, Texas grew rapidly over the period, a fact that tends to correlate with decreasing literary representation relative to population, but it grew less quickly than California or Utah, both of which managed modest increases relative to their multiplying populations. When we examine the details of Texan locations, however, the answer is clear. Before the Civil War, there existed a meaningful subset of fiction interested in the Mexican-American War of 1846–48 (Jeremiah Clemens's *Bernard Lile* [1856] and *Mustang Gray* [1858], Harry Hazel's *The Flying Artillerist* [1853]); after 1861, this subset disappeared and no other comparable event emerged to take its place. War was also a factor in the prominence of New England throughout the period, as Revolutionary War sites such as Lexington and Concord were frequently invoked in both historical and metaphorical contexts.[31] Like war-related Texan locations, however, these too were less frequently invoked after the outbreak of the Civil War, albeit with a less precipitous drop in literary use owing perhaps to the already settled initial rate of reference and to the more enduring national–historic position of the Revolution.

Literary interest in many Confederate states apart from Texas, on the other hand, rose relative to population after the Civil War. These states included Virginia, South Carolina, Louisiana, Tennessee, Georgia, Mississippi, and Alabama, as well as the would-be secessionist Maryland, all of which saw much more direct geographic involvement in the conflict than did Texas. Perhaps most illustrative among these is South Carolina, a moderately large state that flipped sharply from under- to overrepresentation after 1861 almost entirely on the basis of interest in the war, especially via invocations of Charleston and Fort Sumter. Where Sumter, for instance, occurs just once before the war, it is found 158 times thereafter in books published as early as 1861 and with titles—such as Alfreda Eva Bell's *The Rebel Cousins* (1864), Lydia Maria Francis Child's *A Romance of the Republic* (1867), John Esten Cooke's *Wearing of the Gray* (1867), and J. B. Newbrough's *The Fall of Fort Sumter* (an 1867 romance)—indicative of their often sympathetic content despite their publication in the North. More broadly, we observe a southerly shift in the center of literary gravity after the Civil War that does not correspond to any similar human movement (see Figure 5). In short, current events, especially war, appear to be capable of driving relatively rapid shifts in literary attention, a fact that should remind us of the outsize role of topical fiction in the literary marketplace. While we are perhaps

inclined to see literary production as responding primarily to the deep structure of the culture from which it emerges, the swings in literary–geographic attention associated with the events described here suggest that this is not the only force shaping the large body of fiction. The North–South rift, after all, was certainly established long before the beginning of the Civil War (which event is usually said to have secured the subsequent conceptual unity of the nation), yet literary attention seems to have followed the flashpoint of the war itself rather than to have built gradually in response to the formation of that divide.

4. Conclusions

So what did the literary–geographic imagination of mid-nineteenth-century American fiction look like? It was global, certainly, making use of international locations nearly as often as domestic ones. It was also surprisingly and disproportionately urban; although the 20 largest American cities by population made up only about 10% of the national headcount at the time, the 20 most frequently occurring US cities in the corpus accounted for well over a third of all US place-name occurrences. Literary attention was most heavily concentrated along the eastern seaboard, but not especially so in New England and not to the exclusion of the rest of the nation. While literary attention generally lagged the large and growing populations of the Midwest, it did not by any means ignore that region, nor did it overlook the South (especially—but not only—after 1861), nor the West. On the whole, the use of US place names in the fiction of the period correlated reasonably well with population; large places occurred more frequently than small ones to roughly the same degree as the population of the larger location exceeded that of the smaller. But this relationship, while strong, was interestingly imperfect, yielding numerous cases of under- and overrepresentation relative to population.

There exist mixed signals concerning the emergence of American literary regionalism in the years before 1875. The period's strong investment in urban locations suggests that there was at no point a marked preference for the types of rural locales generally associated with the regionalist impulse, nor was there a large-scale shift away from heavily populated regions in the years immediately following the Civil War, when one might have expected to find the early signs of emerging regionalism. Modest changes toward wider distribution of literary attention, especially at the city level, did occur following the war, however, and it remains the case that both before and after 1861 there existed widespread literary use of locations outside the northeast corridor. Whether or not these facts point toward an earlier or later emergence of regional writing—or indeed toward any regionalist

flowering at all—remains an open question in the absence of a broader historical extension of the current research. But they make it almost impossible to maintain that any large-scale shift toward literary regionalism as we conventionally understand it took place in the decade following the war, or that attention to regional locations wasn't already well established by the 1850s.

The notion of the American Renaissance as a phenomenon rooted primarily in New England is also largely unsupported by the data. While New England locations were overrepresented relative to the population of the region both before and after the Civil War, the extent of their overrepresentation actually increased after 1861, a trend that's difficult to reconcile with standard periodizations derived from Matthiessen, which associate the phenomenon with the first half of the 1850s. At the same time, the fraction of all US location uses that fell within New England was hardly overwhelming at around 15%, a figure that indicates the breadth and depth of literary investment elsewhere in the nation and world at the time. Among our most compelling current problems, then, should be to explain this phenomenon both in canonical texts and, pressingly, in texts that more fully represent the shape of the period's literary output.

Finally, the literary–geographic imagination of the period was surprisingly stable over time. True, there were small overall shifts toward greater diversity of locations used after the Civil War and away, on a percentage basis, from some of the largest cities, but these and other changes were on the order of single percentage points in most cases. They were potentially important, but they were not overwhelmingly large. This fact doesn't necessarily suggest that significant shifts weren't taking place over the 25 years in question; indeed it's hard to imagine that the Civil War didn't result in meaningful cultural reconfigurations that are traceable through the period's literary output. But it does suggest that at least in the literary–geographic cases studied here, intellectual significance and the absolute magnitude of the observed effect may be best measured on separate scales; although we're not accustomed to thinking in such terms, a few percent change in a low-level feature measured across thousands of texts might produce tremendously important shifts in the way those texts operate within their cultural field.

I suspect that this incongruity of magnitudes will be something with which we'll often need to grapple as the objects of our literary–historical analysis move to the corpus level. Our existing sense of the changes involved in periodizing transitions is based on close examinations of a modest number of generally very different books, and the examinations themselves have usually been directed toward accentuating those differences. There's nothing necessarily wrong with working that way, by sketching the poles toward which two or more groups of

texts might have been drawn, but it's a recipe for producing periodizing concepts that are much sharper in principle than is likely to be the case across the whole of an era's literary output. When it becomes possible to assess features of that output en masse, we need to be aware of the fact—not new, of course, but newly prominent—that what we're observing is, even in the easiest case, a superposition of individual points between those poles. The aggregation involved almost inevitably damps the magnitude of the observed shift.

It's especially difficult to assess, at this early stage in the development of macro-scale work in literary history, what constitutes an importantly large shift in the observed features of a corpus. I've tried, in the foregoing discussion, to consider the weight that we ought to attach to relatively modest shifts in the distribution of literary–geographic attention, and have concluded that they call into question two major areas of apparently established scholarship. But it's also my hope that the results presented here will provide a useful background against which to evaluate future large-scale critical work using related methods. In any case, corpus-level research isn't the only area that will benefit from the availability of high-level descriptions of literary–geographic use. Conventional literary scholarship will also be enhanced by the ability to assess the degree to which any particular text or cluster of books conforms to or deviates from the trends observed in the larger textual field. It seems useful to know, for instance, just how common it was for a novel written in the 1850s to use locations in New York or California or South Carolina, and to be able to identify shifts in the distribution of geographic attention in which it might participate.

For now, we have the first broadly inclusive survey of American literary–geographic usage in the mid nineteenth century, one that casts light on—and complicates—our longstanding narratives concerning two of the most important periods of American literary history. Increased focus on the international, urban, and slowly evolving nature of the literary–geographic imagination in the US around the Civil War is warranted by the current results, which plot a significant path for future work in both conventional and computationally assisted American literary studies.

Appendix: Technical and Methodological Details

The Corpus

Novels are an especially useful and important aesthetic form through which to assess the development of American literary geography around the Civil War. As Nina Baym has documented, the novel was identified as the preeminent literary form

of its age in the US as early as 1827 (26). By the 1830s, novel production constituted a "deluge" in the words of more than one contemporary observer, one driven largely, Baym argues, by a shift in the composition of the reading public toward the "newly literate masses" (29). Beyond its raw numbers, the novel was said to possess a uniquely direct influence on popular perception due to its intellectual accessibility and social topicality, and to be especially democratic in the population of its authors by dint of its relatively uncomplicated structure (35). By 1850, Baym concludes, there existed a broad consensus on the part of critics, authors, and readers alike that the novel was "*the* literary art form of the nineteenth century," one that represented "the spirit of its age . . . [in] the emergence of the people as a political and cultural force" (44).

To this historical view, we might add Benedict Anderson's famous demonstration that novels provide exactly "the technical means for 're-presenting' the *kind* of imagined community that is the nation" (25).[32] This is to say that novels are and were central to the way in which any nation develops a sense of itself as a socially and historically coherent entity. If we want to understand how Americans saw and represented their nation and its place in the world during the decades surrounding the Civil War, and if we're interested in the relationship between their representations and literary history, we would be hard pressed to find so symptomatic an object as the novel.

The literary corpus used in the present study is thus made up exclusively of American novels and novel-like texts first published between 1851 and 1875. Specifically, it is based on the volumes cataloged by Lyle Wright in his *American Fiction, 1851–1875: A Contribution toward a Bibliography* (1958). Wright's bibliography attempts to list "the fiction . . . written for adults by Americans and printed in the US" between the dates of his title; he specifically excludes reprints, religious tracts, children's literature, genres other than narrative fiction, serials, and books by non-American writers published in the US.[33] Wright consulted both physical copies held in a range of libraries and lists of newly published titles from contemporary sources in the compilation of his bibliography.

Nearly all of the 2,925 titles listed by Wright have been scanned and digitized via optical character recognition (a rapid but error-prone technique) by the Committee on Institutional Cooperation (the current total is 2,774), but only 1,050 of those digital texts have been thoroughly hand-corrected and contain firmly established dates of publication between 1851 and 1875. The present work is based on these 1,050 high-quality, TEI-encoded, datable volumes, which together contain over 80 million words.[34] The research corpus thus comprises 36% of all known American book-form fiction produced during the generation spanning the Civil War. Of these, 489 volumes (containing 36 million words) were published before 1861; 561 volumes (containing 44 million words) were published in 1861 or later. The distribution of volumes by year of publication is shown in Figure 10; the fraction of known volumes represented in the corpus is notably even across the period.

Although the corpus is not large in comparison to those held by Google or HathiTrust, it is of much higher quality than those repositories, includes carefully curated metadata, and excludes the many nonfiction and nonliterary texts that make up the bulk of Google's and Hathi's holdings. These advantages make the present corpus the best currently available for large-scale literary work around the Civil War. It is also the case that the total volume of American fiction produced during the period in question was significant but not immense; as noted above, Wright's bibliography lists fewer than 3,000 volumes published over 25 years.[35] For all of these reasons, it would be difficult to attribute any of the trends or patterns observed in the data to defects in the corpus itself. Nevertheless, it is a long-term goal of the project

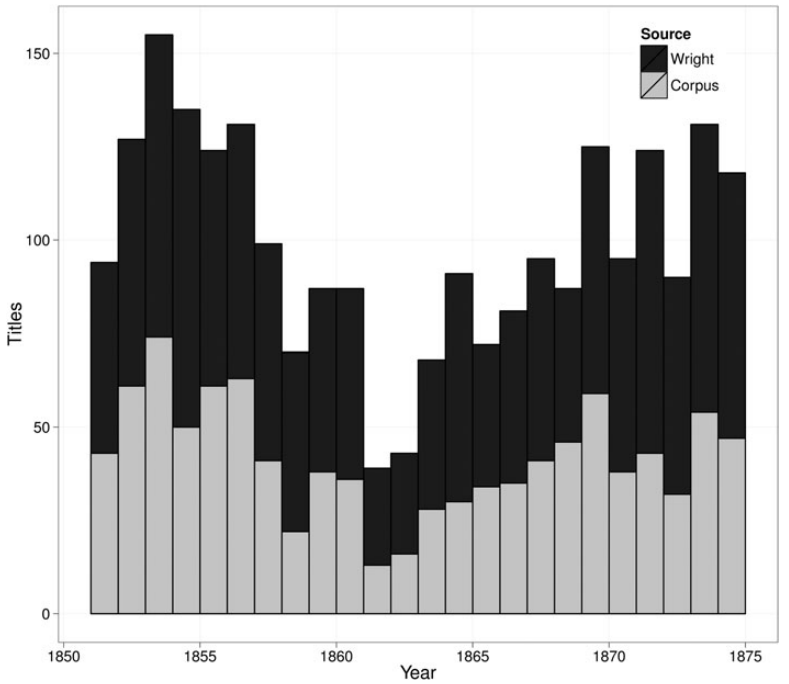

Fig. 10. Number of titles cataloged by Wright (dark gray) and held in the present corpus (light gray) by year of publication. Note that this is not a stacked bar plot; two distinct histograms are here superimposed.

to expand both the depth of its coverage within the period 1851–75 and the breadth of coverage in earlier and later years.

Texts were prepared for analysis first by extracting embedded bibliographic information, then by removing all metadata, apparatus, and XML tags. This produced one plain Unicode text file per volume. These text files in all cases included only the main body of the work, excluding tables of contents, running heads, and other paratext.

Named Entity Extraction

Text strings representing named locations in the corpus were identified using the named entity recognizer of the Stanford CoreNLP package with supplied training data.[36] To reduce errors and to narrow the results for human review, only those named-location strings that occurred at least five times in the corpus and were used by at least two different authors were accepted.[37] The remaining unique strings were reviewed by hand against their context in each source volume. Those that were rarely used as named locations ("Charlotte" is almost always a personal name; "Providence" is used nearly exclusively to mean "divine care") or were hopelessly ambiguous across volumes ("North River," "Mapleton") were discarded. Those that were potentially ambiguous but were used in the corpus predominantly to refer to a single location were annotated to reflect the proper referent. Ancient places were associated with their modern counterparts where possible. After these corrections were applied, there remained 142,706 occurrences of 1,811 unique location strings in the corpus.

A few words on the kinds of locations that are and are not reflected in the data. All place names were extracted from the source texts algorithmically; no

attempt was made to identify or classify by hand the locations those texts evoked, in which the action of their plots took place, or to which they alluded indirectly. For a place to be included—that is, classified as a location by the named entity recognizer—it must have been mentioned by name and used as a proper noun in a source text in a way that indicated it was a location. The algorithm uses both a list of known place names (a gazetteer) and grammatical analysis to recognize locations not previously seen and to exclude as far as possible occurrences of place names not used as such (for example, "Mr. Rochester").

The technique does not attempt to identify or assess the contextual significance of mentioning any individual place name. The significance of a location, for the purposes of this study, attaches to the frequency of its overall usage in the corpus and to variations in that frequency over time (and in response to other data). This approach is obviously different from the ways in which we are accustomed to analyzing space in literary texts. Is the consequent de-emphasis of sentence- and book-level context a problem? The answer can only be that it depends on the questions we're asking. The method is plainly ill-suited to telling us much about any specific occurrence of a named location, at least without further investigation; for that, we need to read the text in which the location occurs. If our questions can only be answered by detailed reference to the function of individual locations in single texts, there is no substitute for close reading. But the alternate approach employed here supplies a different and supplementary kind of context, one that occurs at the level of the corpus (just how unusual is it, say, for a Civil War-era novel to make frequent reference to places in Indiana? Very unusual indeed) and that will prove valuable in its own right. In this larger picture, we would encounter serious difficulties only if texts routinely, systematically, and predominately included place names in which they had little investment and invested heavily in locations they almost never mentioned. This is not the case.

Colloquial names and nonstandard spellings of locations are potential sources of error, but their effect is generally small. Nicknames for cities occur very rarely in the corpus; only two are of even minor note. "Crescent City" for New Orleans occurs 15 times and is mapped to the latter in the process of hand correction. Ten total references to "modern Babylon," generally (but not always) meaning London, are mapped to classical Babylon, hence to Hillah in present-day Iraq. This is less an error than an interpretive choice, one that reads a sentence like "Even the noise and bustle and confusion of London, the great modern Babylon among nations, possessed little or no interest to her" (from James Maitland's *The Cousins*, 1858) as containing one reference to London and one to the ancient city-state. In any case, however, "modern Babylon" occurs just one time in a volume that does not also contain an explicit reference to "London" (and most contain many more than that). Nonstandard spellings are also rare and are generally caught by the grammatical portion of the named entity recognizer, but a small number—often in passages of dialect—do escape detection. Finally, demonyms are excluded on the theory that they refer primarily to persons rather than places. Thus, for instance, the phrase "a Bostonian" contains no locations, but "a woman from Boston" contains one. Like the other potential problem cases, demonyms are relatively rare: "Bostonian" occurs at about 2% of the rate at which Boston does, while "New Yorker" manages less than 0.5% that of New York.

Fictional places present a challenge, though their overall effect is small. Instances in which real places were comprehensively fictionalized in a way that intelligent readers could decode were rare in the period, so represent a minor issue. Place names that were simply invented tend to be tagged as locations by the named entity recognizer, but then produce failed or obviously dubious geocoding results (see the next section) that were excluded by the hand review. That leaves only the class of

ambiguously multireferential names (such as "Mapleton") noted above, a problem not unique to fictional places (most were discarded).

Geospatial Information

The location strings extracted from the corpus texts were associated with geospatial information via Google's geocoding API. This process classified each named location by type ("country," "locality," "natural feature," and so on), standardized its presentation, and associated it with latitude and longitude coordinates or boundaries as appropriate. Geocoding results were further reviewed and a small number of errors corrected. Such errors were often the result of Google's modern dataset and (otherwise desirable, given the subject matter) US bias; where "Naples" occurs in the corpus, for example, it never refers to the present-day city in Florida.

Precision and recall measures were calculated on a small sample of the data. (Precision is the fraction of identified places that are correct; recall is the fraction of actual places in the source text that are identified correctly.) With the above corrections, precision was 0.81, while recall was 0.65; F_1 (an average of precision and recall that provides a unified measure of overall accuracy) was 0.72. Using a looser, more relevant standard that evaluates location occurrences in sum rather than individually, precision rises to 0.86, recall to 0.71, and F_1 to 0.78. These are good results, given the complexity of the problem and the state of the art.[38] Nevertheless, we must bear in mind that the method is not perfect. The hand correction process errs on the side of caution, meaning that most of the locations identified are correct (precision is high), but some locations that do exist in the corpus are not counted in the results (recall is lower). Broadly speaking, we ought not to attach particular significance to any one location occurrence in the absence of a closer examination of its source context, but we may speak with much more confidence about the overall patterns and distributions of location occurrences collectively.

Notes

1. To cite only a handful of recent examples, on literary geography, see Paul Giles, *The Global Remapping of American Literature* (2011); Hsuan Hsu, *Geography and the Production of Space in Nineteenth-Century American Literature* (2010); and (with a complicated relationship to time) Wai Chee Dimock, *Through Other Continents: American Literature across Deep Time* (2006); on public and domestic space, see Christopher Reed, *Bloomsbury Rooms: Modernism, Subculture, and Domesticity* (2004); Thomas Heise, *Urban Underworlds: A Geography of Twentieth-Century American Literature and Culture* (2010); and Miles Orvell and Jeffrey L. Meikle, eds., *Public Space and the Ideology of Place in American Culture* (2009); on the postmodern relationship of time and space, of course, Fredric Jameson, *Postmodernism, or The Cultural Logic of Late Capitalism* (1991) and "The End of Temporality," *Critical Inquiry* 29.4 (2003): 695–718.

2. For the more or less canonical sources, see Matthiessen's *American Renaissance*, Judith Fetterley and Marjorie Pryse's *American Women Regionalists* (1992), and Richard Brodhead's *Cultures of Letters*: *Scenes of Reading and Writing in Nineteenth-Century America* (1993).

3. Consider, in a different vein, the characteristic complaint of Edward Eggleston, who wrote in the preface to his *Hoosier School-Master* (1871) that "it has been in

my mind since I was a Hoosier boy to do something toward describing life in the back-country districts of the Western States. It used to be a matter of no little jealousy with us, I remember, that the manners, customs, thoughts, and feelings of New England country people filled so large a place in books, while our life, not less interesting, not less romantic, and certainly not less filled with humorous and grotesque material, had no place in literature. It was as though we were shut out of good society" (iii).

4. See *American Renaissance* 229n1 for Matthiessen's catalog of the nation's great novelists down to his day. Beyond Hawthorne, he credits New England with only Harriet Beecher Stowe—whose *Uncle Tom's Cabin*, despite being the best-selling American novel of the nineteenth century and having been published in 1852, squarely in the short period he identified as the great flowering of American literature, nevertheless does not merit additional mention in *American Renaissance*—Edward Bellamy, and "minor talents of great distinction" such as Sarah Orne Jewett and Mary Wilkins Freeman.

5. The underlying claim concerning an achieved national unity following the Civil War is also reflected in the long-observed rise in treatment of the US as grammatically singular ("the US is . . ." rather than "US are . . .") after 1865. See G. H. Emerson, "The Making of a Nation," *Universalist Quarterly and General Review* Jan. 1891: 49–67; and Basil Gildersleeve, *Hellas and Hesperia: Or, The Vitality of Greek Studies in America* (1909), who made this argument as early as 1891 and 1909, respectively.

6. See again Fetterley and Pryse.

7. The 1870 census contains known undercounting errors and is therefore excluded from Figure 5; see J. David Hacker, "Recounting the Dead," *New York Times*, New York Times Co., 20 Sept. 2011, Web. US population center data supplied by the US Census Bureau. For an overview of state-level population changes across the period, see Figure 9.

8. All historical population data derived from the US Census Bureau with adjustments for modern geographic boundaries. A similar analysis at the international and city levels is excluded here due to the vastly increased number of regions for which reliable and consistently tabulated historical population figures would be necessary.

9. To illustrate the problem, consider two numbers, say 10 and 100,000. At a later point, these values change to 20 and 110,000, respectively. Which one has undergone the larger (or more significant) change? The answer depends, of course, on whether we care about the relative or the absolute magnitude of that change.

10. For a full description of the method, see Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics* 19.1 (1993): 61–74. For a more humanistically oriented explanation, see Ben Schmidt, "Comparing Corpuses by Word Use," *Sapping Attention* 6 Oct. 2011, Web.

11. Average population for the period was calculated as the mean of geographically corrected census figures for 1850 and 1880. States and territories not included in the 1850 census but included in 1880 were treated as having zero population in 1850. States or territories lacking a population figure in 1880 were excluded from the calculation.

12. 1850 census figures were used for pre-1861 population comparisons; the 1880 census was used for post-1861. Any state or territory not included in the 1850 census was excluded from the calculation.

13. It would be useful to have a comparable figure for twentieth-century fiction to assess more fully the extent of the later shift away from Christian and classical references. Unfortunately, no analogous research is currently possible due to copyright restriction on texts published after 1923. But Google Books data suggest that the frequency of related words such as Israel, Jerusalem, Damascus, Egypt, and Greece all declined markedly (by as much as 80%) in American books between the early nineteenth century and the 1930s (after which Israel and Jerusalem rebound, for obvious geopolitical reasons). See books.google.com/ngrams. For an analysis of the role played by the classics in nineteenth-century America, see Carl J. Richard, *The Golden Age of the Classics in America: Greece, Rome, and the Antebellum US* (2009).

14. Hawthorne's included texts are *House of the Seven Gables* (1851), *The Blithedale Romance* (1852), *The Snow-Image* (1852), *Mosses from an Old Manse* (second edition, 1854), *The Marble Faun*, and *Septimus Felton* (1872). Only *Blithedale*, *Mosses*, and *Marble Faun* contain references to locations of antiquity. *The Scarlet Letter*—not included in the corpus because it was originally published in 1846—contains a single occurrence of "Israel."

15. *Israel Potter* is not a source of error in this case.

16. See Wilkens, "Contemporary Fiction by the Numbers," *Post45* 1 (2011): n. pag., Web. To be clear, "diversely outward looking" is a description of a geographic distribution, not a moral evaluation; the suggestion isn't that the texts of the period necessarily display any special broadmindedness or appreciation of cultural diversity as we now understand those terms.

17. This is even more true, of course, if we consider the various Western territories as lying outside the boundaries of the US before their accession to statehood.

18. On reprinting and its relationship to the publishing trade in nineteenth-century America, see Adrian Johns, *Piracy: The Intellectual Property Wars from Gutenberg to Gates* (2010), chapters 8 and 11; and Meredith McGill, *American Literature and the Culture of Reprinting, 1834–1853* (2003). Note that reprints of foreign texts are excluded from the present corpus.

19. The rise in prominence of the US in later fiction (both American and global) is supported by Google Books data, which show a sharp rise in the frequency of "United States" around the turn of the last century and long-range declines in the major European nations (with the exception of Germany during the two world wars).

20. By way of illustration, consider the relatively low-frequency geographic phrases "son(s) of the South" and "son(s) of Africa," which occur 14 and 13 times in the corpus, respectively. Both are used exclusively—and with predominately positive connotations—to refer to characters living in the southern US. Both phrases are clearly "about" Americans and American issues. But the former refers, of course, exclusively to whites, while the latter is reserved for American blacks, who are thereby characterized as foreign even in a phrase that is much less bitingly racist than many that appear in the fiction of the period.

21. New England is defined in the conventional way to include the states of Maine, New Hampshire, Vermont, Massachusetts, Connecticut, and Rhode Island. In the present case, references to the frequently employed regional toponym ("New England"), which account for over 16% of total regional occurrences, are also counted.

22. New England experienced substantial population growth during the period (to 40.1 million residents in 1880 from 27.3 million in 1850), but the nation as a whole grew much faster, more than doubling its population over those three decades.

23. For comparison, New York City experienced a similar pattern of population growth coupled with a falling rate of literary representation, though to a lesser degree on both counts. As we will see later, literary representation in general appears to have a difficult time keeping up with rapid demographic changes.

24. US Census Bureau, "Urban and Rural Populations" (CPH-2, 5). Urban areas are legally incorporated areas having a population of 2,500 or greater. Due to the rapid increase in overall US population, the total number of people living in cities was growing much faster than the fraction of the population living there. Between 1850 and 1880, the urban population of the US nearly quadrupled. Urbanization is in many ways the remarkable demographic story of the nineteenth century; in the century between 1820 and 1920, the US moved from over 90% rural inhabitants to more than half urban.

25. Nor did the length of the average volume change in any significant way, for reasons that were probably as much economic and logistical as they were formal.

26. The calculation isn't entirely straightforward. There were 1,048 unique locations used in the corpus between 1851 and 1860. That number rose to 1,150 between 1861 and 1875. But there were also more volumes published in the later period, 564 to 491. New unique locations don't appear at a constant rate with each new volume, because the preceding volumes will already have used many of the same places; the more preceding volumes are in the collection, the more rarely will new unique places appear. If we assume, conservatively, that the quantity of unique locations in a given number of volumes grows at an exponentially decreasing rate (specifically, that this number follows the cumulative distribution function of the normalized standard exponential distribution), we would expect, based on the antebellum numbers, the 564 postbellum volumes to contain 1,132 unique locations. The 1,150 that actually occur aren't much more than that (they're the amount we would expect to occur in 581 volumes rather than the existing 564), but this modest excess does conform to our hypothesis that the range of locations used after the war is broader on a per-volume basis than that used in the preceding decade.

27. See Margaret Jewett Bailey, *The Grains, or, Passages in the Life of Ruth Rover* (1854) and Abigail Scott Duniway, *Captain Gray's Company, or, Crossing the Plains and Living in Oregon* (1859), for instance.

28. Richmond came closest, ranking in the mid-20s throughout the period.

29. At the time of the 1860 census, Orleans parish had 174,000 residents; Cook county, 145,000.

30. Total occurrence counts in the corpus are 235 for Chicago (of which 147 appeared after 1870) and 1,579 for New Orleans (of which 209 appeared after 1870). For additional figures and data, see Wilkens, "Cities, Population Growth, and Literary Attention," *Scalable Reading* 10 July 2012, Web.

31. Interest in transcendentalism also drove a modest portion of the literary investment in Concord, but this effect was small; just 19 of 141 mentions of the town occurred in volumes that also contained "Emerson," "Thoreau," or forms of "Transcendental."

32. Readers of Anderson will recall that *Imagined Communities* posits the newspaper alongside the novel as a form especially suited to nation formation. But to include newspapers here would both weaken the literary relevance of the project and skew the results toward contemporary events in a way that is not desirable. Note two points in particular: (1) The object under consideration is, properly speaking, geographic *desire* as represented in and to the nation, meaning that the (relative) freedom to choose "unimportant" locations is an important one. (2) Novels benefited from national distribution in ways that newspapers generally did not, meaning that novels could and did imagine a national readership unlike that of any newspaper.

33. For a full statement of Wright's selection principles, see the preface to volume 3 of his study, where he writes: "Included are novels, short stories, tall tales, allegories, tract-like tales, and fictitious biographies and travels. I have intended to omit . . . publications of the American Tract Society, Sunday school unions, and religious denominations; collections of anecdotes; juveniles, jestbooks, folklore, essays, and periodicals, including subscription series, which were classed as periodicals" *American Fiction, 1876–1900: A Contribution toward a Bibliography* (1966) (ix). Note that works by foreign-born authors were included in the (apparently rare) cases where those authors "considered the US as their permanent home or wrote with coauthors who were American citizens."

34. Digital files were supplied by the Indiana University Digital Library Program, whose support is acknowledged with gratitude. A full list of included titles is available at http://mattwilkens.com/data.

35. For comparison, a conservative estimate of contemporary US fiction production is well over 1,000 volumes per week (Bowker, *Publishing Market Shows Steady Title Growth in 2011 Fueled Largely by Self-Publishing Sector* [2012]; Wilkens, "How Many New Novels Are Published Each Year?" *Work Product* 14 Oct. 2009, Web).

36. See Jenny Rose Finkel, Trond Grenager, and Christopher Manning, "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling," *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* (2005), 363–70.

37. The 5/2 threshold also almost certainly causes a slight overestimation of the fraction of named location occurrences that refer to urban areas, since small towns and rural areas are much more likely to fall below its bar. This effect should be minimal, as the threshold is very low indeed given the size of the corpus.

38. A word on the strict and loose evaluations of accuracy. In the strict case, each hand-identified occurrence of a named location must match its automatically

identified counterpart if the latter is to count as a true positive (or, correct) result. In the loose evaluation, the total numbers of hand- and automatically identified occurrences are compared; if Boston occurs 10 times in each, 10 true positives are recorded, even if some individual occurrences were misidentified (in hypothetically offsetting numbers) by the algorithm. The loose standard is the more appropriate one in the present case, where location counts are treated collectively. No matter which accuracy standard is used, note that the task is particularly challenging because it relies on two stages of automated information extraction. Strings in unstructured natural language text must first be identified as indicators of geographic place, then those strings must be associated with specific current or historical locations. For an evaluation of current toponym resolution methods, see J. L. Leidner, *Toponym Resolution: A First Large-Scale Comparative Evaluation*, School of Informatics: Institute for Communicating and Collaborative Systems (July 2006), Web, PDF (which notes that even human annotators agree with one another concerning the proper identification of named locations only 80%–90% of the time). Both precision and recall could likely be improved for future work by retraining the NER package on period-specific texts.

Works Cited

Anderson, Benedict. *Imagined Communities*. 1983. New York: Verso, 1991.

Baym, Nina. *Novel, Readers, and Reviewers: Responses to Fiction in Antebellum America*. Ithaca: Cornell UP, 1984.

Bercovitch, Sacvan. *The Puritan Origins of the American Self*. New Haven: Yale UP, 1975.

Curtis, George William. *The Potiphar Papers*. New York: Putnam, 1853.

Emerson, Ralph Waldo. "The American Scholar." 1837. *The Collected Works of Ralph Waldo Emerson*. Ed. Robert Ernest Spiller et al. Vol. 1. Cambridge: Harvard UP, 1971. 49–70.

Kirkland, Caroline M. *The Evening Book, or, Fireside Talk on Morals and Manners, with Sketches of Western Life*. New York: Scribner, 1852.

Newell, R. H. *The Walking Doll, or, The Asters and Disasters of Society*. New York: Felt, 1872.

US Census Bureau. Population Division. "Selected Historical Decennial Census Population and Housing Counts." 1990. Web.

Whitman, Walt. "Preface 1855, *Leaves of Grass*, First Edition." *Leaves of Grass and Other Writings*. Ed. Michael Moon. 2nd ed. New York: Norton, 2002. 616–36.

Wright, Lyle Henry. *American Fiction, 1851–1875: A Contribution toward a Bibliography*. 2nd ed. San Marino: Huntington Library, 1965.