

ENVIRONMENTAL STUDIES

Two centuries of settlement and urban development in the United States

Stefan Leyk^{1,2,3*}, Johannes H. Uhl^{1,2}, Dylan S. Connor⁴, Anna E. Braswell^{3,5}, Nathan Mietkiewicz^{3,5}, Jennifer K. Balch^{1,3,5}, Myron Gutmann^{2,6}

Over the past 200 years, the population of the United States grew more than 40-fold. The resulting development of the built environment has had a profound impact on the regional economic, demographic, and environmental structure of North America. Unfortunately, constraints on data availability limit opportunities to study long-term development patterns and how population growth relates to land-use change. Using hundreds of millions of property records, we undertake the finest-resolution analysis to date, in space and time, of urbanization patterns from 1810 to 2015. Temporally consistent metrics reveal distinct long-term urban development patterns characterizing processes such as settlement expansion and densification at fine granularity. Furthermore, we demonstrate that these settlement measures are robust proxies for population throughout the record and thus potential surrogates for estimating population changes at fine scales. These new insights and data vastly expand opportunities to study land use, population change, and urbanization over the past two centuries.

INTRODUCTION

The population of the United States grew from an estimated 5.3 million in 1800 to 309 million people in 2010 (1). On the basis of the definitions from the Census Bureau, the share of the U.S. population living in urban areas grew from 6 to 81% over this period. Urbanization occurred through population growth and the transformation of physical landscapes and ecological systems into developed land. Thus, researchers typically measure these changes through either population- or land-based methods [e.g., (2–6)]. While these two perspectives paint different but complementary pictures of urbanization, they are also sensitive to the scale of measurement. Thus, because of the absence of consistent and detailed, historical information on local land use and local population change, our knowledge of the historical development of the United States is far from complete. Advancing such knowledge would greatly improve our understanding of the broad impacts of urbanization and allow for refined projections of demographics and the built environment.

The absence of detailed historical population data before the mid-20th century severely constrains any population-based assessment of urban processes. Although the U.S. Census records are made publicly available after a period of 72 years, spatially registering and encoding these data are resource-intensive. While researchers have begun to transcribe and extract these data for fine-scale analysis [e.g., (7–9)], publicly available historical population data are accessible only at coarse spatial resolution [e.g., county boundaries; (10)]. This coarse resolution in combination with boundary changes over time (fig. S1) poses a major barrier to studying historical urbanization in the United States using census data [e.g., (11, 12)].

Studying urbanization from a land perspective typically includes land use or land cover data, or, more recently, settlement layers that provide consistent spatial data on the timing, location, and nature

of land use. Although many historical maps contain detailed information on land use over long time periods, their extraction at fine resolution is prohibitively costly because of the volume, complexity, and low quality of such graphical documents (13). Most prior efforts to characterize historical fine-grained settlement or land cover changes rely on remote sensing imagery, which are constrained to the post-1970 era of satellite technology [e.g., (14–17)]. Such historical satellite-derived data are usually coarsely classified, provide limited information on the specific characteristics of built-up land, and are often less accurate for rural areas (18, 19).

In this study, we present a new means of understanding the speed, spread, and nature of urbanization in the United States from 1810 to 2015. We use gridded settlement layers from the Historical Settlement Data Compilation for the United States [HISDAC-US; (20)], which is derived from property records compiled in the Zillow Transaction and Assessment Dataset (ZTRAX). HISDAC-US describes the built environment of most of the conterminous United States back to 1810 at fine temporal (5 years) and spatial (250 m) granularity using different settlement measures. These measures include the number of built-up property records, which can refer to individual properties or units within built-up properties (BUILD) in a grid cell in a given year and the built-up intensity (BUI), or the sum of gross indoor area of all built-up properties. We also extracted for each grid cell the first built-up year (FBUY), which is the earliest construction year on record. For larger analytical units, such as counties, we derived the built-up area (BUA), or the number of grid cells overlapping with one or more built-up properties in a given year.

The principal goal of this analysis is to foreground the value of these novel data in providing insight into long-term settlement and urban development. Building on Leyk and Uhl (20) and other ongoing efforts (9), we leverage the HISDAC-US data to undertake an unprecedented multiscale analysis of the history of U.S. urban development and settlement. These new data can be leveraged to explore and characterize fundamental processes of urban growth through measurement of changes in the built environment, potentially providing insights into the fundamental drivers of development patterns. We anticipate that these measures and insights will provide vast new opportunities to study and understand the history of U.S. urbanization from a land-based perspective.

Copyright © 2020
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
NonCommercial
License 4.0 (CC BY-NC).

¹Department of Geography, University of Colorado Boulder, 260 UCB, Boulder, CO 80309, USA. ²Institute of Behavioral Science, University of Colorado Boulder, 483 UCB, Boulder, CO 80309, USA. ³Earth Lab, University of Colorado Boulder, 4001 Discovery Drive Suite S348, 611 UCB, Boulder, CO 80309, USA. ⁴School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85281, USA. ⁵Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, 216 UCB, Boulder, CO 80309, USA. ⁶Department of History, University of Colorado Boulder, 234 UCB, Boulder, CO 80309, USA.

*Corresponding author. Email: stefan.leyk@colorado.edu

In addition to new land-based assessments of urban change and development, these novel data also unlock new opportunities to model the spatial distribution of population in the past. Our motivation in this regard is rooted in recent research. First, while recent work shows relationships between historical population counts and built-up property attributes, this analysis is confined to the national scale and lacks the spatial detail necessary for understanding variation and change (21). Second, data on developed or built-up land are often used as the main ancillary variable in population modeling using dasymetric refinement approaches. This refinement method is a form of areal interpolation that makes use of relationships between the target variable (population) and the ancillary variables used for subunit estimation [e.g., (22–25)]. Third, parcel data combined with population and road network data were used in recent efforts to study long-term urbanization processes within U.S. cities (26), but such data have not been available for the entire nation to date. Last, researchers applied similar principles of land availability and suitability to disaggregated historical census summary statistics and created fine-resolution population distributions (27). However, these approaches lack robust testing and validation for the years before 2000. Given this body of research, we argue that data products such as the HISDAC-US provide unique opportunities to model not only changes in the built environment but also, potentially, fine-grained historical population estimates. Progress in this area could unlock new possibilities for the spatiotemporal analysis of urbanization in the United States, combining both land- and population-based perspectives.

While our main findings confirm well-known broad diffusion patterns of urbanization, HISDAC-US settlement layers enable us to identify detailed building trajectories as well as expansion and densification patterns at various spatial scales. The fine granularity of the data is depicted by maps of the FBUI (Fig. 1) and the number of built-up property records (BUILD; Fig. 2, B to F) from 1810 to 2010. Finer-scale data break down broader national and regional development trends and describe, for example, local processes of expanding urban and suburban areas or infilling in built-up places during different time periods. We also demonstrate relationships

between settlement measures and population growth. We estimate that, on average, each additional built-up property at the county-level is associated with around 2 to 2.25 additional people with some regional variation. This finding is notable as there are, at present, no reliable estimates of long-term, fine-resolution population growth for the United States. Thus, we argue and demonstrate that the novel HISDAC-US data provide an unprecedented opportunity to study and understand long-term urbanization and settlement processes at fine spatial and temporal granularity from the beginning of the 19th century to today.

RESULTS

Taking a land perspective on urbanization: Where, when, and how much land was built-up?

Using the fine temporal and spatial granularity as well as different built characteristics, we elucidate new spatiotemporal settlement patterns. With these patterns, we draw a detailed picture of the evolution of built-up land use in the conterminous United States from 1810 to 2015.

Mapping the earliest recorded built-up properties (FBUI) within boundaries of varying spatial scale, we find that urban development trends are strongly dependent on the size of the spatial unit used (broad national to local; Fig. 1). By using the contemporary county boundaries of the 2010 decennial census as consistent mapping units, we observe two primary, well-known national trends (Fig. 1, A and B). First, we find trends of urban development diffusing westward from Northeastern and coastal Southern states into the interior of the United States, including the eastern parts of Texas, Kansas, and Arkansas. These trends unfold later in the Appalachian Mountains and parts of Florida, likely because of the rough topography and limited habitability of these areas. Second, while buildings in counties in the central and western states of the United States tend to be newer than their Eastern counterparts, there are isolated counties across the western seaboard and interior regions (e.g., Denver and Wichita Falls; Fig. 1, D and E) that experienced particularly early waves of

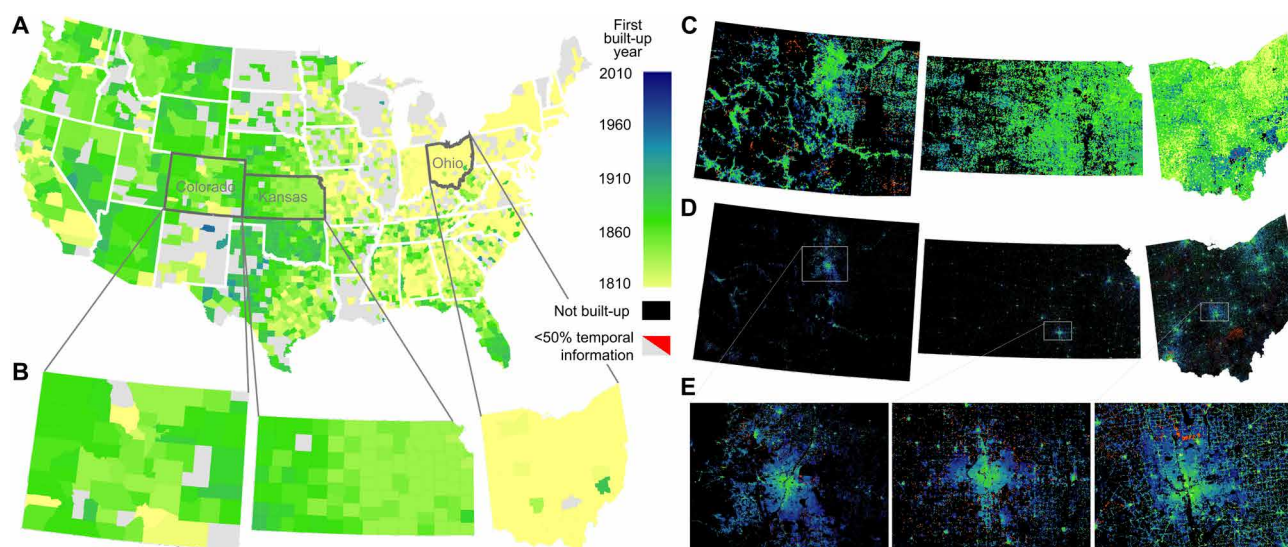


Fig. 1. Maps of the FBUI at different levels of granularity. The maps depict national-, regional-, and local-scale processes of human development: (A) county-level FBUI within contemporary (Census 2010) boundaries used as constant units of analysis over time (counties where no built-up year is available are shown in gray); more detailed distributions of FBUI for the states of Colorado, Kansas, and Ohio within (B) county boundaries, (C) 2500-m grid cells, and (D) 250-m grid cells, respectively. (E) A detailed depiction of the 250-m resolution data for the cities of Denver, CO; Wichita Falls, KS; and Columbus, OH.

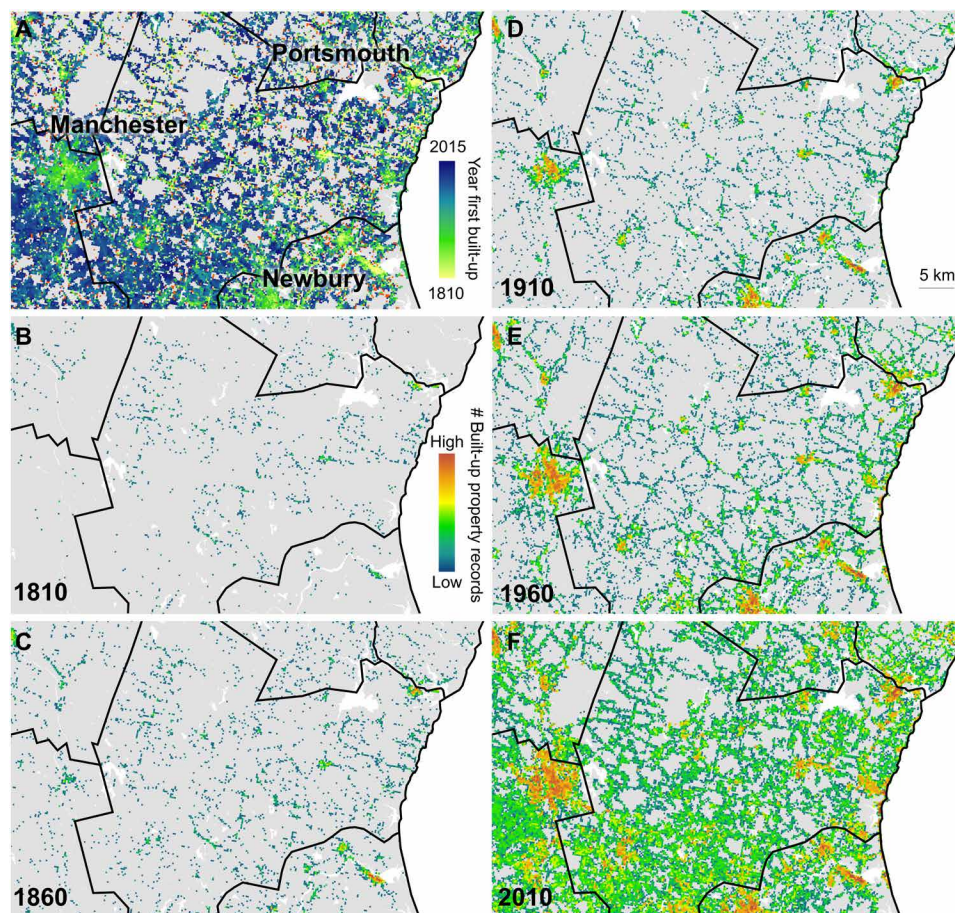


Fig. 2. HISDAC-US settlement layers derived from the ZTRAX data at fine granularity for different points in time. The layers are shown at 250-m spatial resolution for different points in time, 1810–2010, for Rockingham County, NH and surroundings, including: (A) FBUIY layer in which raster cells are assigned the earliest built year recorded, and a time series of the number of built-up property records (BUILD) located within a raster cell in (B) 1810, (C) 1860, (D) 1910, (E) 1960, and (F) 2010. County boundaries of the 2010 census are shown in black.

development relative to the rest of their respective states. In many instances, these nodes of early development predate the demarcation of these regions as U.S. states.

By assigning FBUIY values to smaller spatial units such as individual grid cells of specific size (e.g., 2,500-m resolution, Fig. 1C; or 250-m resolution, Figs. 1D and 2A), we are able to assess local settlement trends within consistent spatial units that break down the county-level patterns. For example, early settlement and growth along Colorado's Front Range emanates from a number of isolated centers, with Denver being the largest (Fig. 1, D and E, left). Also, earliest records of development in more rural settings of the state (in mountainous areas or in the plains) appear spatially related to streams and topographic conditions that facilitated development, livelihood, access to water, and transportation. In contrast, new development in Kansas spreads as a broader national pattern of westward expansion, rather than as sprawl from discrete larger urban hubs (Fig. 1, D and E, middle). In Ohio, urban centers begin to overlap over time as they expand into one another (Fig. 1, D and E, right). These patterns illustrate the opportunities provided by such multiresolution data for detecting local-to-regional scale settlement and land development trends over long time periods.

We used the built year of each property in combination with building attributes to compute time series of various settlement variables at

different resolutions to more holistically measure local and regional development trends. As discussed above, these settlement measures include the total number of built-up properties (BUILD), the BUI of land derived from the sum of indoor floor area of existing built-up properties, and the number of grid cells built-up within a chosen unit or BUA (see Materials and Methods for details). We extracted these variables within consistent spatial units across time periods to generate long-term trajectories (e.g., fig. S2 at the state-level) and multitemporal spatial distributions (e.g., fig. S3 at the county-level) to characterize variation in settlement patterns over time. We illustrate county-level estimates of BUILD and its change every 5 years between 1810 and 2015 and spatial clusters for each point in time (movie S1). However, the full details of local settlement processes can only be uncovered at the finest granularity.

Taking Rockingham County, NH and the areas surrounding New Hampshire and Massachusetts as an example region, we trace spatial distributions of BUILD at the finest spatial resolution of 250 m over five points in time (1810, 1860, 1910, 1960, and 2010; Fig. 2, B to F). This analysis allows us to track the number of built-up property records at the grid cell-level and better understand local urban growth processes. In this particular case, the cities of Manchester, NH; Newburyport, MA; Amesbury, MA; and Portsmouth, NH grew as separate small urban hubs until 1860. Fewer built-up properties

were established in rural parts of the area along roads during the early and mid-1800s. By 1910, Manchester grew substantially, in area and density, while the port cities developed at slower rates. This trend continued, and by 1960, low-density settlement in rural areas had expanded along roads to increasingly connect higher-density urban hubs. Furthermore, during this time period, development increased rapidly along the coastline. Last, by 2010, the area had experienced intensified sprawl in its southern and coastal regions, a continued expansion of urban hubs, and increasing densification in the South, which grew into a larger urban and suburban conglomerate. Such subcounty, temporal settlement patterns have the potential to yield vast new insight into the geographical unfolding and intensity of local urban development processes.

Land-based measures of change characterize types of urban development at varying scales

Fine-scale settlement layers provide unique opportunities to distinguish between land-based processes of urbanization such as expansion and densification. Expansion refers to the amount (or proportion) of new developed area over time, and densification is the ratio of the change in BUI to the change in BUA over time (see details in Materials and Methods).

Coarser-scale, county-level maps of expansion and densification reveal notable regional variation (see fig. S4 and movie S1 for a complete sequence of those maps and their spatial cluster maps). These results complement the observed regional settlement patterns but provide more details about the underlying processes of urban growth, often a function of time, infrastructure, and access to technology. Maps of peak timing of densification and expansion (fig. S5) reveal that both processes are temporally associated and vary regionally. For example, along the coastlines of the Southeast and the Southwest of the United States, the vast majority of counties have expansion peaks earlier than densification, indicative of land expansion maxima followed by maximum infilling in already built-up areas. We found the opposite process in the noncoastal Northeast, the Midwest, and parts of the Mountain West. In these areas, development and peak densification occurred over the early to mid-1900s, and expansion—often in the form of sprawl—subsequently unfolded and peaked during the second half of the 20th century.

We assessed city-level measures of expansion and densification for San Francisco, CA; Atlanta, GA; and Boston, MA (Fig. 3A). With the exception of the time period from 1920 to 1950, which was a period of rapid rise and decline in terms of expansion and densification, in Boston, both measures trended gradually upward over time but took opposite trends after 2000 (expansion declining and densification rising). Atlanta and San Francisco exhibit more notable variation. In the sprawling city of Atlanta, densification has remained modest (with some recent increases), but expansion has markedly increased since the mid-20th century. Over the past decade, Atlanta had decreasing expansion and increasing densification. For San Francisco, in contrast, we find the opposite pattern: Expansion remained relatively low over time, but densification continued to rise sharply. These trends are consistent with the widely held view of Atlanta as a sprawling metropolitan region and the greater compactness of San Francisco.

From these trajectories, we can track the development of a city at fine temporal resolution over 200 years and visualize accompanying spatial change patterns at the grid cell-level. By assessing the change in BUA (ΔBUA ; i.e., locations that were developed during a given time period) and the change in BUI (ΔBUI ; i.e., interior area added per grid cell during a given time period) in detailed maps (Fig. 3,

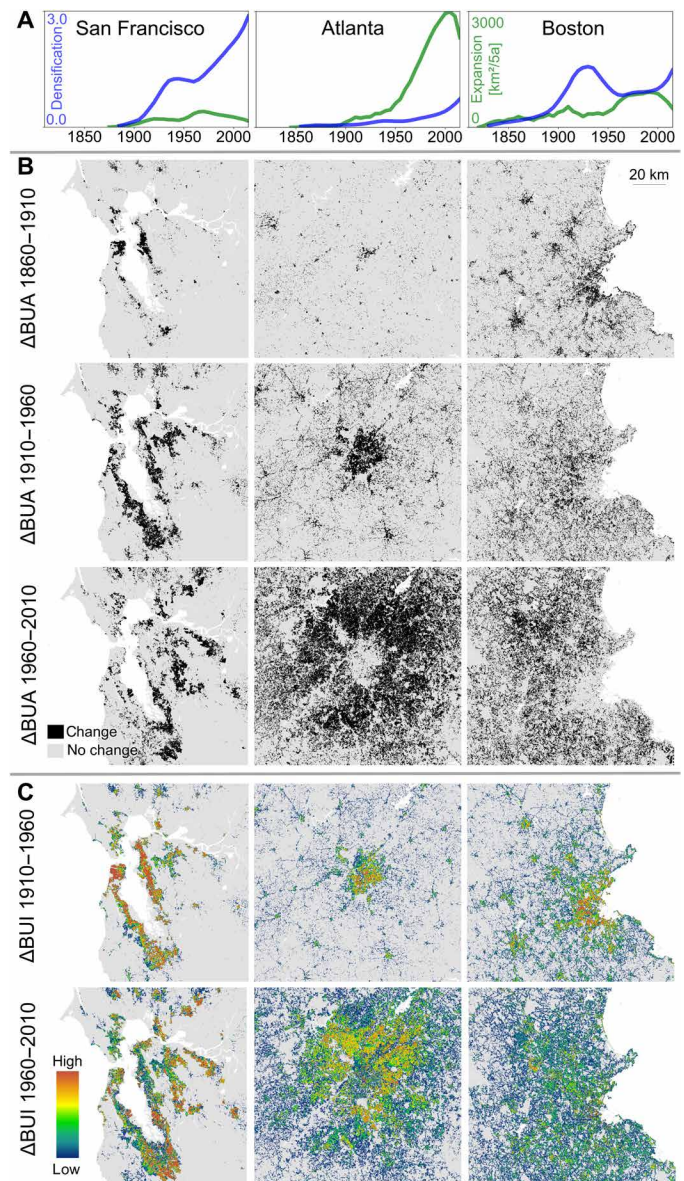


Fig. 3. Settlement trends and multitemporal distributions for San Francisco, Atlanta, and Boston. (A) Time series of densification and expansion, calculated over 15-year time increments computed within metropolitan statistical area boundaries of 2010. (B) New built-up grid cells (indicated by black grid cells) during the given time periods. (C) Change in BUI (i.e., the sum of building indoor area per grid cell) during the given time periods, with warmer colors indicating greater change.

B and C, respectively), we gain insight into the development mechanisms generating differences across cities. The spatial patterns suggest that San Francisco (Fig. 3B, left) developed under topographic constraints allowing limited new development and creating notable changes in density over the past 100 years. Atlanta (Figs. 3B, middle), in contrast, had a massive increase in developed area since the 1960s and developed into one of the most sprawling cities in the United States with low building density. Last, the spatial patterns for Boston (Fig. 3B, right) are illustrative of a city with an early-developing and high-density urban core. Continued new development and increases in density were more balanced in Boston over the past century than in the other two cities. Thus, these novel data products provide vast

new opportunities for measuring and testing proximate patterns and determinants of urban spatial development (e.g., topographical influences on land use). To further illustrate the dynamism of these data products, we developed fine-grained distributions of BUI with a temporal resolution of 5 years for these three cities, as well as Los Angeles, CA; Dallas–Fort Worth, TX; and Philadelphia, PA between 1810 and 2015 (movie S2).

Dissecting and measuring forms of growth at fine scale in urban and rural areas

Across the conterminous United States, we find that land-based expansion and densification show converging and diverging trends over recent decades, particularly in more developed counties. We created trends of densification (Fig. 4A) and absolute expansion (Fig. 4B) for counties in two strata, which we refer to as rural and urban, over time. The rural stratum is composed of counties that have less than the 66th percentile of BUI across all counties (using the 2010 census boundaries), calculated individually for each year. The urban

stratum is defined by counties with BUI values greater than the 66th percentile. This stratification allows us to assess how settlement in relatively more and less developed places changed over time.

The two strata have different trajectories for both measures with significantly higher values in the urban stratum. For urban counties, both measures have an increasing trend up to the 1930s (Fig. 4, A and B). After 1940, expansion increases markedly until the early 2000s but decreases notably during the past decade (Fig. 4B). Densification has varying trends since 1930: It levels off for a short time, increases between 1940 and 1960, then decreases until the 1980s, and since then, increases sharply until 2010 (Fig. 4A). The rural stratum shows continuous increases in both measures, steepest for densification between 1910 and 1960 and for expansion between the 1940s and 1980s, somewhat temporally offset to densification. Both measures remained relatively constant between 1980 and 2010. Counties in the urban stratum have significant variability indicating wide ranges of expansion and densification values, likely found in different regions. In general, we find compelling differences in comparing the two

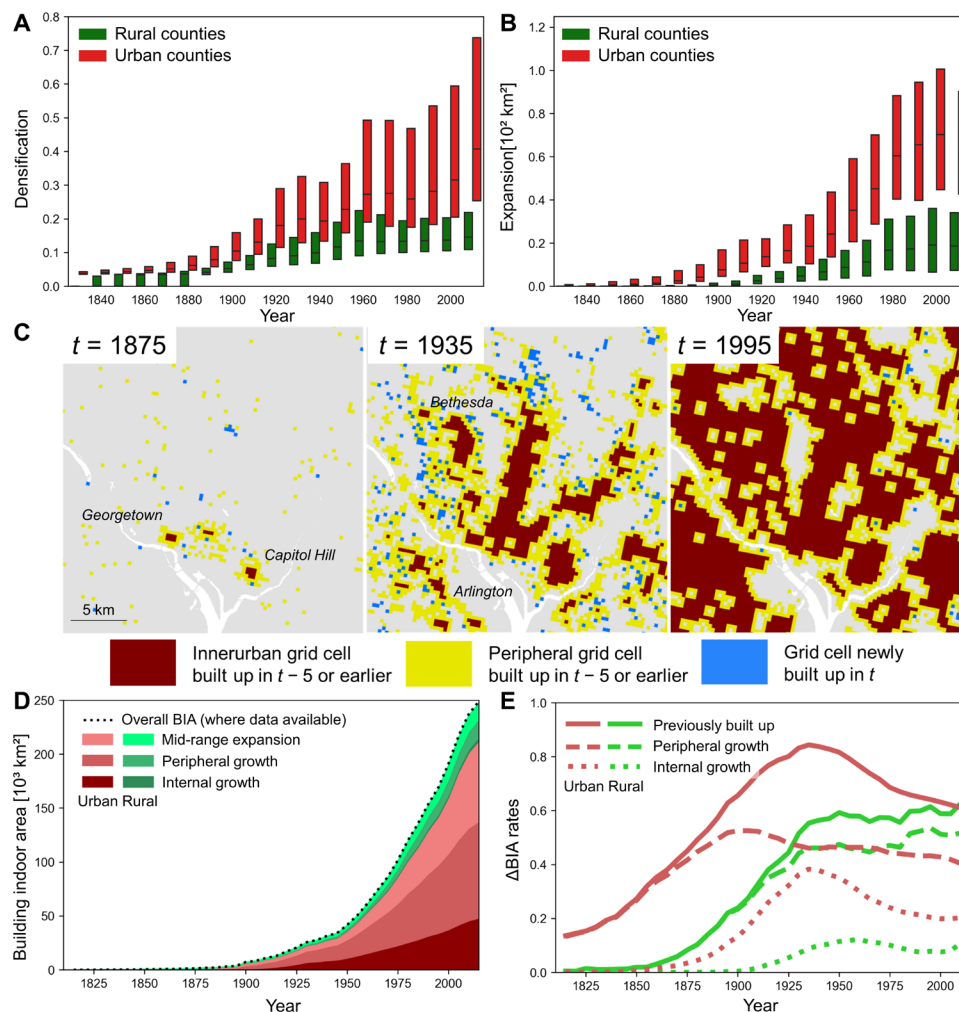


Fig. 4. Settlement trends describing different types of growth in rural and urban strata. Boxplots of semi-decadal distributions of (A) densification and (B) absolute expansion in rural and urban counties. (C) Graphic display of relative locations of newly and previously built-up grid cells to calculate midrange expansion, internal, and peripheral growth within the Greater Washington, DC area including Arlington, Bethesda, and Georgetown. (D) Trends of building indoor area (BIA) in urban and rural strata (counties), each broken down into whether the increase happened in newly built-up cells (midrange expansion), in previously built-up cells at the edge of larger BUAs (peripheral growth), or in previously built-up cells in inner parts of BUAs (internal growth). (E) Proportion of internal, peripheral, and internal-peripheral combined growth (i.e., growth in previously built-up cells) in relation to overall change for the two strata (rural and urban).

strata that appear to characterize the rural-urban divide in the development of the conterminous United States.

To better understand the observed growth patterns, we spatially decomposed the trends of built-up interior area (BIA), which is the BUI aggregated across the whole United States, within both rural and urban counties into different types of growth. The different growth categories include midrange expansion (i.e., the appearance of newly built-up cells), peripheral growth (i.e., in previously built-up cells at the edge of larger BUAs), and internal growth (i.e., in previously built-up cells in inner parts of BUAs; Fig. 4C). The resulting trends illustrate the magnitudes of BIA across and within strata (Fig. 4D). By 2010, BIA in the urban stratum is roughly 10 times greater than in the rural counties. Within the urban stratum, the dominant type of growth has been peripheral growth followed by midrange expansion. These two types of growth are very similar in the rural stratum. Internal growth has the lowest values of BIA, but its proportion has been notably higher in the urban stratum in the past.

To examine the relationships and changes between different types of growth across each stratum, we computed ratios of changes of BIA in previously built-up cells (i.e., peripheral and internal growth) to all changes in BIA (in previously and newly built-up cells; Fig. 4E). For urban counties, we see a steep increase of the proportion of previously built-up land to overall growth until a peak in the early 1930s, when approximately 85% of new growth happened as either peripheral or internal growth. This percentage declined to approximately 62% in 2010, likely as a result of increased expansion (newly built-up land, often in the form of sprawl). Peripheral growth, which is higher than that of internal growth, has a peak around 1900 at 55% and since declined to 42%. In contrast, internal growth increased steeply until it reached a peak in the early 1930s at 40% and declined until 2000 to 25%. During the past decade, internal growth shows a slight uptick, which corresponds to increasing densification, seen in urban counties. In rural counties, the proportion of internal and peripheral growth combined increases steeply to approximately 60% in the 1950s and since then shows varying trends between 55 and 65%. As internal growth never exceeded 20%, most of these trends are driven by peripheral growth.

The main trends in rural and urban counties converge over time, indicating that by 2010, the proportion of growth in previously built-up cells to growth in newly built-up cells is very similar in both strata (between 60 and 62%). This convergence also indicates that during the past seven decades, the proportion of growth due to expansion has been increasing in urban counties and slightly decreasing in rural counties. We expect to find significant regional variability in these patterns if evaluated for different geographic units (e.g., states or counties), describing deviating trajectories for different criteria used for defining urban and rural strata.

Supporting a population perspective of urban development: Settlement as a reliable predictor of historical and contemporary population

We conclude our results by using a panel analysis approach to illustrate that built characteristics can meaningfully capture human settlement and urbanization patterns (28). This method serves as a test for whether the settlement layers can support population modeling for the study of urban development. In this analysis, we predicted population counts from the decennial censuses of 1860, 1910, 1960, and 2010 by the number of built-up property records (BUILD) observed in the HISDAC-US data; we tested all land-use types together

and residential land use only. We relied on BUILD because it has the highest overall correlation with population counts over time in comparison to other settlement measures (Fig. 5). Through this analysis, we attempted to accomplish two objectives. First, we examined how much of the temporal variation in population can be explained using BUILD. Second, we estimated the number of people associated with each additional built-up property in a county.

On the basis of the R^2 values for a pooled ordinary least squares (OLS) regression model of all counties from 1860 (Table 1), BUILD based on all land-use types explains almost 93% of the variation in population across counties over time (column 1; $R^2 = 0.926$). This result holds even when we restrict the sample to counties with consistent boundaries through time (<10% change in area measures; column 2; $R^2 = 0.898$; see Materials and Methods for details). As these models include no other control variables, we conclude that BUILD appears to be highly effective in characterizing county-level changes in population over time. We suspect that much of the variation across these models is a function of changes in household size and the distribution of dwelling units by size over time and space. The estimates from the standard OLS models suggest that, on average, one unit increase in BUILD is associated with an increase of around 2.6 to 2.7 people. There are, however, many difficult-to-observe reasons for why counties with more or less built-up properties differ in population (e.g., many coastal cities have both economic opportunity and high-density building stock due to land constraints).

We also ensure the robustness of our results to omitted variables by presenting more conservative estimates when regressing changes in population on changes in BUILD (columns 3 to 5 of Table 1). We ran a least squares dummy variable (LSDV) model with county-level fixed effects (column 3) in which the model variation comes from population changes within counties over time [within estimator; (29)], revealing a consistent and significant relationship of around 2.2 people for each additional built-up property within a county. We ran a generalized least squares (GLS) estimator (column 4) and controlled for potential decadal trends in population and BUILD (column 5), producing generally consistent estimates. Thus, our analyses suggest that, on average, an additional built-up property in each county is roughly associated with a 2.2- to 2.25-person increase in the total population. Although the quality of HISDAC-US data is considerably poorer before 1860, our analyses using earlier starting points yield very similar results (table S1). Results were very similar for BUILD based on all land-use types ($\beta = 2.246$, $R^2 = 0.873$), as well as residential land-use types only ($\beta = 2.246$, $R^2 = 0.875$). We examined the effect of regional variation by running the same GLS estimator shown in column 5 of Table 1 for the four regions Northeast, South, Midwest, and West (table S2). Coefficients vary between 2.029 and 2.537, indicating low levels of regional variation in the statistical relationship at the county-level. While these results are robust and provide strong indication of predictive power of BUILD for population at the county-level, the observed effects of spatial and temporal variability have to be further investigated, particularly at finer spatial scales.

DISCUSSION

Fine-scale spatial and temporal data improve our understanding of long-term settlement patterns

Settlement patterns can only be fully understood from a multiscale perspective (30) that characterizes local, regional, and national patterns of urban development and land-use change. Through our unique

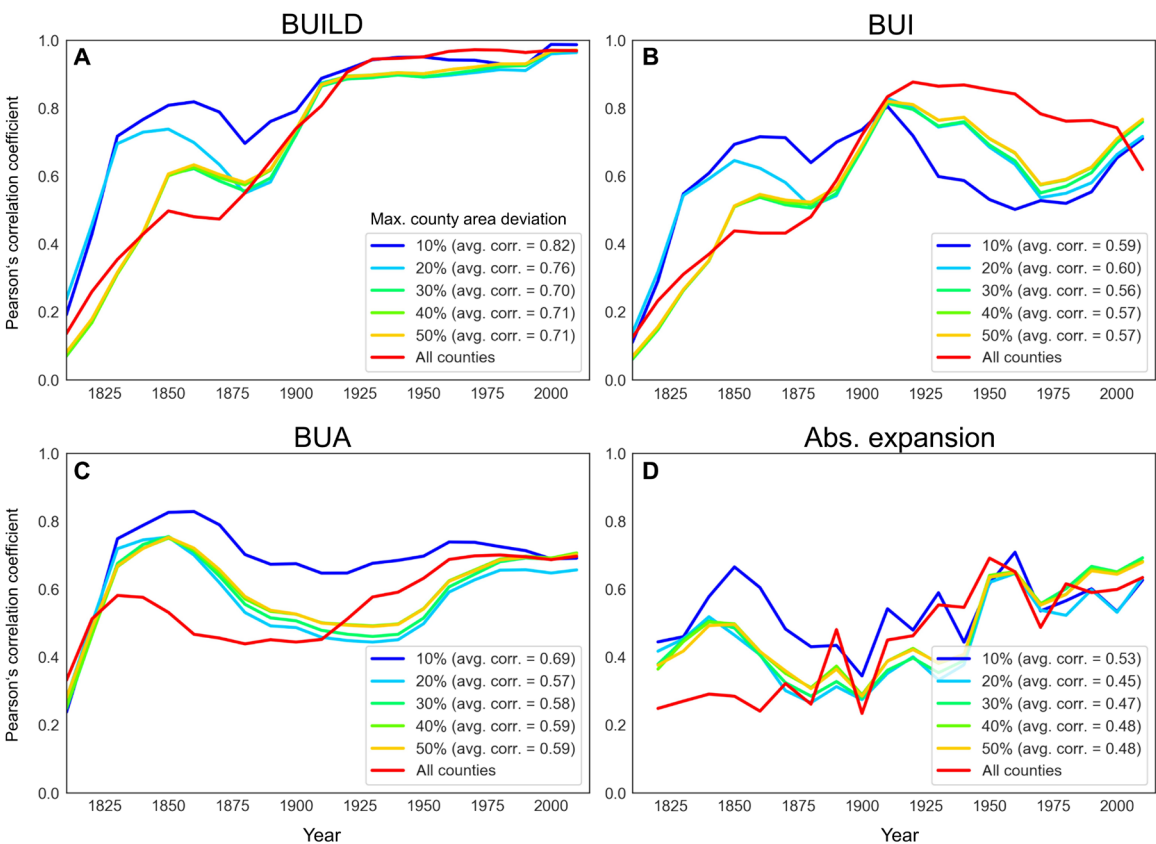


Fig. 5. Correlation measures between county-level population counts and settlement variables. The different plots show correlations over the time period from 1810 to 2010 between population and (A) the number of built-up property records (BUILD), (B) BUI, and (C) BUA. avg. corr., average correlation. In (D), correlation measures between county-level population change and absolute expansion (Abs. expansion) are shown for the same time period. Population counts are enumerated within historical county boundaries, while settlement measures (all land use classes) are calculated within contemporary county boundaries. Correlation measures are shown for various levels of temporal county boundary stability (e.g., the blue lines represent only counties whose area did not change more than 10% over time) to demonstrate the importance of compatible spatial units in spatiotemporal analysis. Average correlation coefficients are shown over all years in parentheses.

Table 1. Panel analysis results using different points in time, 1860–2010. SEs are given in parentheses. Statistical significance is provided (* $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$). The different models tested are OLS, least squares dummy variables (LSDVs), and generalized least squares (GLS). LSDV and GLS are “within” estimators. BUILD is based on all land-use types; results for residential land-use types only are very similar.

	OLS	OLS	LSDV	GLS	GLS
	(1)	(2)	(3)	(4)	(5)
	Total population	Total population	Total population	Total population	Total population
Number of built-up property records (BUILD)	2.680***	2.668***	2.190***	2.190***	2.246***
SE	(0.007)	(0.012)	(0.058)	(0.013)	(0.014)
Constant	10503.8***	13767.2***	22513.3***	22513.3***	16123.0***
SE	(550.069)	(815.994)	(1068.970)	(701.136)	(1274.106)
N	10996	5900	5900	5900	5900
R ²	0.926	0.898	0.946	0.864	0.873
Adjusted R ²	0.926	0.898	0.929	0.819	0.831
County sample	All	Consistent	Consistent	Consistent	Consistent
Fixed effects	–	–	County	County	County, decade
Clustered SE	No	No	Yes	Yes	Yes

data products with unprecedented temporal coverage and fine spatial and temporal resolution, we are able to provide new multiscale depictions of historical settlement. From these depictions, we identify time periods of slow or fast growth and characterize different urban processes that are only discoverable at very fine scales. Our results document and analytically evaluate regional and local patterns depicting rural-urban transformations, urban expansion and peripheral growth, as well as densification and infilling processes.

We envision that our new measures on when structures were built (FBUY), the number of built-up property records (BUILD), the BUI, and the BUA at a given point in time as well as derived process measures such as expansion or densification will enable new opportunities to answer scientifically and theoretically grounded questions in urban research. For example, in ongoing projects, we have started to deploy these data to better understand the changes in the built environment that unfolded in U.S. cities related to residential segregation, postwar suburbanization, and the more recent resurgence of central city areas (9). Thus, we see enormous potential in these data for examining landscape evolution, fragmentation, and the role of technology, economic, and social forces in shifting the contours of urban development.

Methodologically, the use of gridded settlement time series allows researchers to conduct their analyses consistently with studies that apply remote sensing images [e.g., (31, 32)] and extract urban or developed land within any spatial unit. However, the HISDAC-US data are less limited temporally than remote sensing products that cover time periods of no more than three to four decades. Furthermore, the HISDAC-US layers are more accurate (20), richer in attribution related to the built environment, and cover a time period of more than 150 years for most of the conterminous United States. By tackling critical process questions in urbanization, we use the new HISDAC-US data-derived measures of development to connect data-scientific analysis of large spatiotemporal data and substantive inquiry in urban geography, demography, and land use science.

Detailed built environment attributes enable holistic examination of settlement and urban development

Temporal trajectories of different settlement measures within spatial units of interest, such as counties, cities, or tracts, provide a detailed picture of the complexity of long-term development in the United States. This knowledge of development fuels our understanding of when, where, and how quickly humans have urbanized the country. Evaluating the interrelationship between different settlement measures is essential to understanding how the nature of urban development has differed across time and space. We demonstrate that settlement is difficult to describe in either univariate or linear terms, and different development attributes follow timelines that vary across urban strata and regions, which are likely dictated by existing infrastructure, technology, and the developability of land (such as in coastal ecosystems). Complementing other findings [e.g., (33)], we also demonstrate that processes such as densification and expansion are interrelated temporally. However, we found that the synchrony between the peaks of those processes varies greatly across regions and cities, pointing to different forms of historical settlement and urban development. These types of development vary markedly between coastal and interior areas, northern and southern regions, and with topographic constraints and environmental conditions. With these insights, researchers can draw an unprecedented picture of the nature and timing of rural, suburban, and urban development in the United States

at varying scales. The advances in our understanding of settlement processes have the potential to inform ongoing discussions about the spread and compactness of urban areas (2, 34, 35).

Following the paradigm of “people are where people build,” this study demonstrates an effective way to estimate historical population at fine spatiotemporal granularity

There is a common understanding in the fields of rural studies, urban geography, and demography that the built environment is related to population and other demographic attributes (21, 27). These insights provide the basis for a population-based perspective on urban development assessments. Our panel analysis results demonstrate that historical settlement layers in HISDAC-US (20) are associated with population at relatively fine spatial granularity (i.e., counties). Such results are important in two distinct but related ways.

First, the predictive power of the population models indicates that the settlement-population relationship is highly robust. These models enable us to build county-level population data over more than 150 years at fine temporal resolution. These data help overcome the dependence on traditional decadal census surveys [e.g., (36)] and may support the creation of future population assessments to improve population projections. Such model outcomes can be used to create time series of consistent population estimates (e.g., within contemporary county boundaries from the 2010 census) to perform unprecedented temporal analysis. Using these analytical innovations, demographers and urban modelers can study demographic processes related to rural-urban transitions over long time periods at meaningful spatial scales and inform population projections.

Second, the robust settlement-population linkages indicate the potential for reproducing such population models within different spatial units including census units of finer spatial granularity (e.g., census tract boundaries of 2010) or alternative geographic units. For example, researchers might need to estimate population and its changes within certain land cover classes or zones of high vulnerability to natural or industrial hazards. The fine resolution of the settlement layers makes it possible to model population at fine scales using attributes such as the number of built-up property records or BUI allocated to such alternative analytical units. Such advances will greatly benefit research on coupled socio-environmental systems and improve our understanding of existing interrelationships and processes. However, variance in the relationship between population and built environment attributes across time and space requires further investigation. While our comparison of county-level relationships by region produces quite consistent results, we have yet to investigate the stationarity of these relationships at finer spatial scales. We suspect that land-use type will play a particularly crucial role in inferring small-area population quantities from built-environment data.

Novel and extensive spatial information necessitates serious investigation into uncertainty and potential data limitations

While the use of such novel data layers opens unprecedented research opportunities, it is important to instruct and educate the data user on existing uncertainties. In Leyk and Uhl (20), some of these temporal, positional, and thematic uncertainties are reported, assessed, and measured in detail. For our analysis, the settlement layers were systematically corrected on the basis of focal raster operations and adjusted using census data (see Materials and Methods). However, while these adjustments reduce some of the inherent bias and result in population models with high predictive power, the reported missingness in the

original ZTRAX data will still potentially cause underestimation of settlement and has to be considered for critical use of the data products in subsequent analyses. We expect these issues to further improve as Zillow continues to update their database, but certain data gaps will always remain. Furthermore, it is important to note that temporal information, such as the FBUI, does not necessarily indicate the year of the first settlement but represents the earliest built years on record in the ZTRAX database of currently existing buildings. Thus, we may not know about earlier built units that have been demolished and rebuilt (or not rebuilt) or still exist but miss built-year records (Fig. 1B). This uncertainty varies across regions and can be addressed by sensitivity analysis and detailed case studies where high-quality data are available.

Future opportunities

Future steps to leverage these new opportunities will explore the creation of settlement estimates at finer spatial and temporal resolution as well as the inclusion of demographic variables and ancillary data for alternative geographies to fully use the potential of these data layers for improved fine-scale urban and population modeling. Of particular interest is the estimation of alternative demographic and housing-related attributes to create a more insightful picture of the human-built environment and its population. These fundamental components will enable the research community to advance research and theory on urban studies, land use science, natural hazards, landscape ecology, and other interdisciplinary pursuits (9, 37, 38). Using the attribute richness of the settlement data, researchers can explore questions of great societal importance at spatial and temporal scales relevant to the operational scale of urban and human-environmental processes including local rural-urban transitions, changes in ecological services, and trends in land fragmentation.

MATERIALS AND METHODS

Settlement and census data

We use the ZTRAX to derive data products that can be used for the extraction of settlement measures at different points in time. ZTRAX is a geocoded housing and property-level database based on existing cadastral data sources that contains more than 374 million data records for approximately 200 million parcels in over 3100 counties in the United States (<https://zillow.com/ztrax>). Zillow Group is an online real estate database company that was founded in 2006. We extracted attributes such as the land-use class, the construction year of the structure on a parcel, and geolocation information (e.g., an approximate location for an address point) to create time series of raster layers. The workflow for creating the spatiotemporal database model, an SQLite database with spatial query extension, and the data products used in this study are described in full detail in Leyk and Uhl (20). The data layers are collected in the HISDAC-US, which is organized as a collection of datasets at the Harvard Dataverse repository (<https://dataverse.harvard.edu/dataverse/hisdacus>). First, we produced a series of semi-decadal raster layers representing the BUI, the sum of gross indoor area of all built-up properties in a grid cell (250 m by 250 m) in a given year between 1810 and 2010. Second, for the same time period, we also produced a series of semi-decadal raster layers representing the number of built-up property records (BUILD) in a cell in a given year. Third, we built a composite raster layer that indicates for each raster cell the first year a built unit was established (FBUI). Last, we derived the BUA as the number of grid

cells in a spatial unit of interest (e.g., counties) with at least one built unit in a given year. The spatial resolution of all raster layers is 250 m, and the temporal resolution available in HISDAC-US is 5 years.

HISDAC-US also contains uncertainty layers at the pixel and county levels (20) that the data user is urged to use for the assessment of positional, temporal, and thematic uncertainty. First, there are proportions of records without a construction year in some counties. Also, in some instances, the year refers to the most recently built unit, and it remains unknown whether there has been a structure before; in other cases, there are several built years given, indicating the very first year and the most recent one, for example. Second, the land-use class attributes vary across counties and states but have been generalized and consolidated to some degree, making them more comparable across the nation. Third, the latitude/longitude records are missing for a portion of the records prohibiting fine-scale localization of the records but indicate the county. The geolocation records represent approximations for the corresponding address, and thus, there is inherent positional uncertainty that needs to be addressed.

Census data and boundary files at the county-level were collected from the National Historical Geographical Information System [NHGIS; (10)]. We used the contemporary county boundaries (2010 census) to extract ZTRAX measures at different points in time. To build our population models at the county-level, we used nominal population statistics (persons count) in 1810, 1860, 1910, 1960, and 2010 and the corresponding time-specific county boundaries from the NHGIS website (fig. S1) as well as the number of housing units in 2010 for our correction procedure, as described below.

Data correction and geoprocessing

We extracted the settlement measures (BUILD, BUI, and BUA) from the raster time series within contemporary county boundaries (2010 census) using zonal statistics geoprocessing functions to create settlement measures for different points in time within consistent spatial units. To mitigate some of the data quality issues, particularly the missingness of built-year records as described above, we applied a spatiotemporal correction procedure to improve county-level settlement measures at different points in time as follows. We carried out this procedure for all variables using built-up properties of all land-use types together and for the BUILD variable based on residential land-use type only to test both corrected data versions in the population model.

We assumed that records in the database without a built year exist at present if they indicate the presence of a built-up property (i.e., in 2015, which is the most recent year in the currently available ZTRAX database) and the likelihood of the actual built year is the same across all years.

For each county, we computed the proportion of missing built-year records (TMiss) in 2015

$$TMiss = \text{SumBYMiss} / (\text{SumBuilt2015} + \text{SumBYMiss}) \quad (1)$$

where SumBYMiss is the sum of missing built-year records and SumBuilt2015 is the sum of built-up properties in 2015 with built-year records. Depending on the magnitude of TMiss (i.e., $TMiss < 50\%$, $TMiss > 50\%$, $TMiss = 100\%$; these thresholds can vary as needed), we corrected the contemporary and earlier county-level settlement measures. Of the 3108 counties in 2010, 1636 counties had less than 25% TMiss; 2201 counties had less than 50% TMiss. The spatial and statistical distributions of county-level TMiss are shown in fig. S6.

First, for counties with $T_{\text{Miss}} < 50\%$ (or another user-defined threshold), relative changes in BUILD, BUI, and BUA were considered reliable. Thus, assuming that records without built-year information existed in 2015, a corrected $\text{BUILD}_{2015,\text{corr}}$ per county was calculated as

$$\text{BUILD}_{2015,\text{corr}} = \text{BUILD}_{2015,\text{uncorr}} + \text{SumBYMiss} \quad (2)$$

A correction factor was calculated as

$$c_{\text{BUILD}} = \text{BUILD}_{2015,\text{corr}} / \text{BUILD}_{2015,\text{uncorr}} \quad (3)$$

Then, each value of the county-level BUILD time series was multiplied with c_{BUILD} , resulting in a corrected BUILD time series while preserving relative changes between years as observed in the uncorrected data.

To correct BUI, for each county, the average BUI per built-up property in 2015 was calculated as

$$\text{BUI}_{\text{AVG}} = \text{BUI}_{2015,\text{uncorr}} / \text{BUILD}_{2015,\text{uncorr}} \quad (4)$$

and then multiplied with the corrected BUILD value in 2015 resulting in the adjusted county-level BUI in 2015

$$\text{BUI}_{2015,\text{corr}} = \text{BUILD}_{2015,\text{corr}} \times \text{BUI}_{\text{AVG}} \quad (5)$$

In analogy to Eq. 3, a correction factor c_{BUI} was calculated and then applied to the whole BUI time series for each county. The BUA time series layers (with value 1 for grid cells with one or more records that had a built year and value 0 for all other cells) were corrected slightly differently. For each county in 2015, we created another binary layer, BUA_0 , with value 1 for those grid cells that contained at least one record without a built year and value 0 for all other cells. We then calculated the area of the spatial union of BUA_{2015} and BUA_0 , which results in the corrected 2015 BUA

$$\text{BUA}_{2015,\text{corr}} = \text{BUA}_{2015,\text{uncorr}} \vee \text{BUA}_0 \quad (6)$$

Earlier BUA layers were then corrected using a correction factor c_{BUA} , calculated in analogy to c_{BUILD} and c_{BUI} . Second, for counties where $T_{\text{Miss}} > 50\%$, changes in BUILD, BUI, and BUA were not considered reliable. As before, BUILD in 2015 was corrected by SumBYMiss. We then derived relative change estimates in BUILD between different points in time based on the five nearest counties where $T_{\text{Miss}} < 50\%$. These average regional gradients of BUILD were used to retrospectively extrapolate BUILD to earlier points in time. To correct BUI in these unreliable counties, we interpolated the average BUI values per built-up property found in the five nearest counties where $T_{\text{Miss}} < 50\%$, multiplied them with the corrected BUILD values, and extrapolated the resulting BUI values to earlier data layers in the time series while preserving the average relative changes in the five reliable neighboring counties. Similar to the reliable counties above, the BUA in 2015 was corrected by the spatial union of BUA in 2015 and BUA_0 . These corrected values were then extrapolated retrospectively while preserving the relative change between years derived from BUA gradients within neighboring counties where $T_{\text{Miss}} < 50\%$.

Once the above correction steps were finalized, BUILD, BUI, and BUA values were estimated for those counties where there was

no information at all. Using the corrected time series resulting from the steps above, BUILD, BUI, and BUA for each year were interpolated using the corresponding values from the nearest five counties where $T_{\text{Miss}} < 50\%$.

Last, we further adjusted the corrected and extrapolated settlement measures BUILD and BUI using the number of housing units published by the U.S. Census in 2010 as follows. First, for each county in 2010, we used the difference between census housing unit counts and BUILD to adjust BUILD in 2010. Then, we adjusted BUILD for the whole time series while preserving the relative changes between years. Using these adjusted values of BUILD in each year, we adjusted the BUI time series proportionally. The BUA time series could not be corrected using census data, because there is no reference information on the spatial distribution of census housing unit counts within counties and thus no BUA-compatible measure.

Expansion and densification calculation

We used the extracted settlement measures to derive variables that indicate more implicitly the process of change. We calculated relative and absolute expansion as the proportion and absolute value of new developed area, respectively

$$\text{Expansion}_{\text{rel}} = (\text{BUA}_{t1} - \text{BUA}_{t0}) / \text{BUA}_{t0} \quad (7)$$

$$\text{Expansion}_{\text{abs}} = (\text{BUA}_{t1} - \text{BUA}_{t0}) \quad (8)$$

where BUA_{t0} and BUA_{t1} are the BUA estimates for the beginning and ending year, respectively. This measure was used to evaluate the amount of change in developed area over a given number of years, reflecting how much development has been added, absolutely and proportionally to the initial condition, respectively. We also calculated densification, which is the change in BUI over the change in BUA

$$\text{Densification} = (\text{BUI}_{t1} - \text{BUI}_{t0}) / (\text{BUA}_{t1} - \text{BUA}_{t0}) \quad (9)$$

where BUI_{t1} and BUI_{t0} are the built-up intensities for the beginning and ending year of the considered time period, respectively. This measure quantifies the increase in BUI in proportion to newly developed areas over a given number of years.

Statistical analysis

We created the maps of the local indicators of spatial association [LISA; (39)] to identify statistically significant spatial clusters in the county-level distribution of the target variable (e.g., change in BUILD; 999 permutations; $P < 0.05$). A hot spot is a statistically significant high-high (HH) cluster, i.e., a high value that is surrounded by other places with high values to constitute a statistically significant group of counties of higher values. Accordingly, a cold spot [low-low (LL) cluster] indicates a low value surrounded by other low values. Counties labeled with HL and LH represent statistically significant outliers from the spatial distribution.

Panel analysis allows us to control for individual-unit heterogeneity and thus variables that may explain differences across counties (e.g., cultural or architectural differences) unmeasured or variables that change over time but not across counties (time-invariant characteristics such as policies, technological advancement, or regulations). This way, panel analysis makes it possible to detect and measure effects that cannot be observed in either the modeling of cross-sectional data or purely descriptive time series analysis (28). To examine the

relationship between the number of built-up property records (residential and all land use, corrected) as predictor, and population (person counts) as the outcome variable within an entity (county), we used two-way fixed-effects panel models to account for such forms of heterogeneity. Thus, we include county and time period fixed effects to help account for this bias, and assess the net effect of the predictors on the outcome variable by allowing the model intercept to vary across the spatial units as well as over time. The equation for the (time and entity) fixed-effects regression model is

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \dots + \beta_k X_{k,it} + \gamma_2 E_2 + \dots + \gamma_n E_n + \delta_2 T_2 + \dots + \delta_t T_t + u_{it} \quad (10)$$

where Y_{it} is the dependent variable with i = entity and t = time, $X_{k,it}$ are the independent variables with coefficients β_k , u_{it} is the error term, E_n is the county n [$n - 1$ entities included as binary (dummies) in the model] with the coefficient γ_n for the binary regressors (entities), and T_t is the binary variable (dummy) for time (there are $t - 1$ time periods) with coefficient δ_t for the binary time regressors.

We compared OLS-based balanced panels with LSDV- and GLS fixed-effects models to better understand the impact of fixed effects on the estimators' predictive power. We included all counties in the balanced panel that remained sufficiently compatible over time, i.e., counties whose areas do not change more than 10% compared to the contemporary county boundaries over the entire time period. All settlement variables were tested but because of multicollinearity issues, only individual ones could be used at a time. Data extraction, analysis, and statistical modeling have been carried out in Python and STATA; geoprocessing steps have been done using Feature Manipulation Engine (FME) and the ArcGIS 10.6 Arcpy Python package as well as NumPy, Pandas, and Matplotlib.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/6/23/eaba2937/DC1>

REFERENCES AND NOTES

- M. J. Anderson, *The American Census: A Social History* (Yale Univ. Press, 2015).
- D. Balk, S. Leyk, B. Jones, M. R. Montgomery, A. Clark, Understanding urbanization: A study of census and satellite-derived urban classes in the United States, 1990–2010. *PLOS ONE* **13**, e0208487 (2018).
- E. G. Irwin, N. E. Bockstael, The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20672–20677 (2007).
- V. C. Radeloff, R. B. Hammer, S. I. Stewart, Rural and suburban sprawl in the U.S. Midwest from 1940 to 2000 and its relation to forest fragmentation. *Conserv. Biol.* **19**, 793–805 (2005).
- K. C. Seto, M. Fragkias, B. Güneralp, M. K. Reilly, A meta-analysis of global urban land expansion. *PLOS ONE* **6**, e23777 (2011).
- K. C. Seto, J. S. Golden, M. Alberti, B. L. Turner II, Sustainability in an urbanizing planet. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 8935 (2017).
- J. Logan, W. Zhang, in *The Routledge Companion to Spatial History*, I. Gregory, D. DeBats, D. Lafreniere, Eds. (Routledge, London, 2018), chap. 11, pp. 21–151.
- S. E. Spielman, J. R. Logan, Using high-resolution population data to identify neighborhoods and establish their boundaries. *Ann. Assoc. Am. Geogr.* **103**, 67–84 (2013).
- D. Connor, K. Clement, A. Cunningham, M. Gutmann, S. Leyk, How entrenched is the spatial structure of neighborhood inequality? Evidence from the integration of census and housing data for Denver from 1940 to 2016. *Ann. Am. Assoc. Geogr.*, (2019).
- S. Manson, J. Schroeder, D. Van Riper, S. Ruggles, IPUMS National Historical Geographic Information System: Version 12.0 [Database]. *Minneapolis: University of Minnesota*, 39 (2017).
- P. E. Beeson, D. N. DeJong, W. Troesken, Population growth in U.S. counties, 1840–1990. *Reg. Sci. Urban Econ.* **31**, 669–699 (2001).
- V. C. Radeloff, S. I. Stewart, T. J. Hawbaker, U. Gimmi, A. M. Pidgeon, C. H. Flather, R. B. Hammer, D. P. Helmers, Housing growth in and near United States protected areas limits their conservation value. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 940–945 (2010).
- Y.-Y. Chiang, S. Leyk, C. A. Knoblock, A survey of digital map processing techniques. *ACM Comput. Surv.* **47**, 1–44 (2014).
- C. Homer, J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, K. Megown, Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **81**, 345–354 (2015).
- C. Homer, J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. J. McKerrrow, J. N. Van Driel, J. Wickham, Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogramm. Eng. Remote Sens.* **73**, 337–341 (2007).
- J. H. Uhl, S. Leyk, Towards a novel backdating strategy for creating built-up land time series data using contemporary spatial constraints. *Remote Sens. Environ.* **238**, 111197 (2019).
- M. Pesaresi, G. Huadong, X. Blaes, D. Ehrlich, S. Ferri, L. Gueguen, M. Halkia, M. Kauffmann, T. Kemper, L. Lu, M. A. Marin-Herrera, G. K. Ouzounis, M. Scavazzon, P. Soille, V. Syrris, L. Zanchetta, A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE J. Selected Top. Appl. Earth Observ. Remote Sens.* **6**, 2102–2131 (2013).
- J. D. Wickham, S. V. Stehman, L. Gass, J. Dewitz, J. A. Fry, T. G. Wade, Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sens. Environ.* **130**, 294–304 (2013).
- S. Leyk, J. H. Uhl, D. Balk, B. Jones, Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States. *Remote Sens. Environ.* **204**, 898–917 (2018).
- S. Leyk, J. H. Uhl, HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Sci. Data* **5**, 180175 (2018).
- M. C. P. Moura, S. J. Smith, D. B. Belzer, 120 years of U.S. residential housing stock and floor space. *PLOS ONE* **10**, e0134135 (2015).
- B. Semenov-Tian-Shansky, Russia: Territory and population: A perspective on the 1926 Census. *Geogr. Rev.*, 616–640 (1928).
- J. K. Wright, A method of mapping densities of population: with Cape Cod as an example. *Geogr. Rev.* **26**, 103–110 (1936).
- J. Mennis, Dasytmetric mapping for estimating population in small areas. *Geogr. Compass* **3**, 727–745 (2009).
- S. Leyk, A. E. Gaughan, S. B. Adamo, A. de Sherbinin, D. Balk, S. Freire, A. Rose, F. R. Stevens, B. Blankespoor, C. Frye, J. Comenetz, A. Sorichetta, K. M. Manus, L. Pistolesi, M. Levy, A. J. Tatem, M. Pesaresi, The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth Syst. Sci. Data* **11**, 1385–1409 (2019).
- C. Barrington-Leigh, A. Millard-Ball, A century of sprawl in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8244–8249 (2015).
- Y. Fang, J. W. Jawitz, High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Sci. Data* **5**, 180067 (2018).
- B. Baltagi, *Econometric Analysis of Panel Data* (John Wiley & Sons, ed. 4, 2008).
- P. D. Allison, *Fixed Effects Regression Models* (SAGE publications, 2009), vol. 160.
- M. Xu, C. He, Z. Liu, Y. Dou, How did urban land expand in China between 1992 and 2015? A multi-scale landscape analysis. *PLOS ONE* **11**, e0154839 (2016).
- B. Bhatta, S. Saraswati, D. Bandyopadhyay, Urban sprawl measurement from remote sensing data. *Appl. Geogr.* **30**, 731–740 (2010).
- D. H. Nong, C. A. Lepczyk, T. Miura, J. M. Fox, Quantifying urban growth patterns in Hanoi using landscape expansion modes and time series spatial metrics. *PLOS ONE* **13**, e0196940 (2018).
- M. Sapena, L. Á. Ruiz, Analysis of land use/land cover spatio-temporal metrics and population dynamics for urban growth characterization. *Comput. Environ. Urban. Syst.* **73**, 27–39 (2019).
- M. Wolff, D. Haase, A. Haase, Compact or spread? A quantitative spatial model of urban areas in Europe since 1990. *PLOS ONE* **13**, e0192326 (2018).
- D. Haase, N. Kabisch, A. Haase, Endless urban growth? On the mismatch of population, household and urban land area growth and its effects on the urban debate. *PLOS ONE* **8**, e66531 (2013).
- Y. Fang, J. W. Jawitz, The evolution of human population distance to water in the USA from 1790 to 2010. *Nat. Commun.* **10**, 430 (2019).
- J. Uhl, D. Connor, S. Leyk, A. Braswell, Urban spatial development in the United States from 1910 to 2010: A novel data-driven perspective (2020); <https://ssrn.com/abstract=3537768>.
- H. Zoraghein, S. Leyk, Data-enriched interpolation for temporally consistent population compositions. *GISci. Remote Sens.* **56**, 430–461 (2019).
- L. Anselin, Local Indicators of Spatial Association—LISA. *Geogr. Anal.* **27**, 93–115 (1995).
- P. Siczewicz, E. Kelley, J. Long, *U.S. Historical Counties* (Newberry Library, Chicago, Illinois, USA, 2011).

Acknowledgments: The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We acknowledge access to the Zillow

Transaction and Assessment Dataset (ZTRAX) through a data use agreement between the University of Colorado Boulder and Zillow Group Inc. More information on accessing the data can be found at <http://zillow.com/ztrax>. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group. Support by Zillow Group Inc. is acknowledged. We thank three reviewers for their constructive comments. **Funding:** Funding for this work was provided through the Humans, Disasters, and the Built Environment program of the National Science Foundation, Award Number 1924670 to the University of Colorado Boulder, the Institute of Behavioral Science, Earth Lab, the Cooperative Institute for Research in Environmental Sciences, the Grand Challenge Initiative and the Innovative Seed Grant program at the University of Colorado Boulder as well as the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Numbers R21 HD098717 01A1 and P2CHD066613. Publication of this article was co-funded by the University of Colorado Boulder Libraries Open Access Fund. **Author contributions:** S.L., J.H.U., D.S.C., and A.E.B. conceptualized and designed the study and planned the analysis. S.L. and J.H.U. designed the gridded data products and developed the dissemination strategies. J.H.U. processed the data and created the figures and movies. J.H.U., D.S.C., and S.L. performed the statistical analysis. S.L. wrote the paper and secured the data.

J.H.U., D.S.C., A.E.B., J.K.B., M.G., and N.M. revised the paper and provided the substantive inputs on design and result interpretation. S.L., M.G., and J.K.B. secured funding. All authors have contributed significantly to this research and have seen and approved the final version.

Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data layers are collected in the HISDAC-US, which is organized as a collection of datasets at the Harvard Dataverse repository (<https://dataverse.harvard.edu/dataverse/hisdacus>). Additional data related to this paper may be requested from the authors.

Submitted 20 November 2019

Accepted 10 April 2020

Published 3 June 2020

10.1126/sciadv.aba2937

Citation: S. Leyk, J. H. Uhl, D. S. Connor, A. E. Braswell, N. Mietkiewicz, J. K. Balch, M. Gutmann, Two centuries of settlement and urban development in the United States. *Sci. Adv.* **6**, eaba2937 (2020).

Two centuries of settlement and urban development in the United States

Stefan Leyk, Johannes H. Uhl, Dylan S. Connor, Anna E. Braswell, Nathan Mietkiewicz, Jennifer K. Balch and Myron Gutmann

Sci Adv **6** (23), eaba2937.

DOI: 10.1126/sciadv.aba2937

ARTICLE TOOLS

<http://advances.sciencemag.org/content/6/23/eaba2937>

SUPPLEMENTARY MATERIALS

<http://advances.sciencemag.org/content/suppl/2020/06/01/6.23.eaba2937.DC1>

REFERENCES

This article cites 31 articles, 4 of which you can access for free
<http://advances.sciencemag.org/content/6/23/eaba2937#BIBL>

PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2020 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).