Computers, Environment and Urban Systems 78 (2019) 101383

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities

Yuan Lai[a], Constantine E. Kontokosta[a,b,*]

[a] Center for Urban Science & Progress and Department of Civil and Urban Engineering, New York University, United States of America
[b] Marron Institute of Urban Management, New York University, United States of America

ARTICLE INFO

Keywords:
Building alteration
Topic modeling
Machine learning
Big data
Natural language processing

ABSTRACT

Building alteration and redevelopment play a central role in the revitalization of developed cities, where the scarcity of available land limits the construction of new buildings. The adaptive reuse of existing space reflects the underlying socioeconomic dynamics of the city and can be a leading indicator of economic growth and diversification. However, the collective understanding of building alteration patterns is constrained by significant barriers to data accessibility and analysis. We present a data mining and knowledge discovery process for extracting, analyzing, and integrating building permit data for more than 2,500,000 alteration projects from seven major U.S. cities. We utilize natural language processing and topic modeling to discover the thematic structure of construction activities from permit descriptions and merge with other urban data to explore the dynamics of urban change. The knowledge discovery process proceeds in three steps: (1) text mining to identify popular words, popularity change, and their co-appearance likelihood; (2) topic modeling using latent Dirichlet allocation (LDA); and (3) integrating the topic modeling output with building information and ancillary data to discover the spatial, temporal, and thematic patterns of urban redevelopment and regeneration. The results demonstrate a generalizable approach that can be used to analyze unstructured text data extracted from permit records across varying database structures, permit typologies, and local contexts. Our machine learning methodology can assist cities to better monitor building alteration activity, analyze spatiotemporal patterns of redevelopment, and more fully understand the economic, social, and environmental implications of changes to the urban built environment.

1. Introduction

City agencies maintain vast databases of information relevant to city management, urban planning, and infrastructure investment, but these resources are often buried deep in legacy information technology systems, cordoned off from the general public and other agencies by inaccessible data structures and non-standard formats (Bettencourt, 2014; Kontokosta, 2018). Increasing access to these city administrative records can provide new sources of information to understand urban activity patterns and neighborhood dynamics. However, most administrative data represent a "digital exhaust", where the potential uses are often far different than the original rationale for collecting the data (Harford, 2014). To extract new insights from these data sources, efficient and scalable data mining and analytical modeling techniques are needed (Offenhuber & Ratti, 2014).

Within the city administrative structure, the Department of Buildings (DOB) is responsible for the safety and regulation of the built environment.[1] Construction activities reflect the economic, regulatory, and social dynamics of the city and are an important indicator of material consumption, population growth and loss, energy use, and waste generation (Sartori, Bergsdal, Müller, & Brattebø, 2008). Building alterations,[2] in particular, play a central role in the revitalization of developed cities, where the scarcity of available land limits the construction of new buildings. The adaptive re-use of existing space reflects the underlying economic activity of the city and can be a leading indicator of economic development and diversification.

However, it is often challenging to extract large-scale data for research and analysis from a city's building department. Although these

* Corresponding author at: Marron Institute of Urban Management, 60 5th Avenue, 2nd Floor New York, New York University, NY 10011, United States of America.
E-mail addresses: yuan.lai@nyu.edu (Y. Lai), ckontokosta@nyu.edu (C.E. Kontokosta).

[1] Although this city agency is most commonly referred to as the Department of Buildings, there a range of specific titles across different municipalities. Throughout this paper, we refer to the DOB as a generic descriptor of the respective city agency responsible for building and construction activity oversight.

[2] We broadly define the term "alteration" as construction activity that alters an existing building, which can involve addition, renovation, retrofit, and change of use. We elaborate on this definition in the Data section and Table 1.

https://doi.org/10.1016/j.compenvurbsys.2019.101383
Received 2 April 2019; Received in revised form 12 July 2019; Accepted 2 August 2019
Available online 17 August 2019
0198-9715/

departments maintain detailed records of permit applications, alteration work, and violations (among other related activities), these data are collected primarily for code enforcement and regulatory compliance. As such, data are often structured to facilitate data entry and archival preservation, and not collected, organized, or coded for analysis and decision support. These barriers to data sharing across agencies and to integration with other ancillary data create unnecessary roadblocks for the general public and other city agencies.

This study presents a generalizable data mining methodology to extract and analyze building construction permit data at high spatial and temporal resolution. Using natural language processing (NLP) and topic modeling, we develop a scalable approach to understand the thematic structure and spatiotemporal patterns of building alteration activities extracted from building department records. We apply this method to seven U.S. cities (New York City, Boston, Los Angeles, San Francisco, Chicago, Seattle, and Austin) and demonstrate how it can be used to standardize data for analysis across cities and agencies with varying data definitions and structures, as well as localized geographic, political, and land use characteristics.

To the best of our knowledge, there are no existing studies on knowledge discovery in databases (KDD) using detailed construction permit application data. Our goal is to objectively identify building alteration topical themes based on permit descriptions. Our research (1) creates an extensive alteration permit database by integrating records from multiple cities; (2) applies topic modeling to discover the thematic structure of alteration activity; and (3) analyzes the spatiotemporal patterns of building alteration as an indicator of the hyperlocal dynamics of urban development. We first introduce our research motivation followed by data descriptions and ancillary data integration methods. We then describe our methodology that includes text mining, topic modeling, and time-series analysis. A discussion follows of our results for each of the seven cities, and their implications for urban data science and data-driven city planning. The paper concludes with a summary of key findings, methodological limitations, and future work.

2. Research motivation

Construction activities - including new development, demolition, and alteration - have a profound local impact on resource consumption, economic growth, and neighborhood change (Beccali, Cellura, Fontana, Longo, & Mistretta, 2013; De Melo, Goncalves, & Martins, 2011; Helms, 2003; Juan, Gao, & Wang, 2010; Lees, 2003). For instance, retrofitting existing buildings has become a primary method to improve energy efficiency in cities, but the lack of measurement and monitoring makes the nature and scale of this type of construction activity largely unknown (Kontokosta, 2013). Although many cities have digitized building permit applications to enable public dissemination of relevant information (Shadbolt et al., 2012), these data are structured for enforcement and record-keeping, resulting in untapped “data tombs” (Fayyad & Uthurusamy, 2002; Neef, 2014). Moreover, each city manages permit data based on its specific urban and regulatory context, resulting in a range of naming conventions, variable definitions, and included fields. Although this practice may be sufficient for local administrative functions, the lack of standardization constrains information sharing and comparative studies within and across cities (Ku & Gil-Garcia, 2018).

Of all the information contained in construction permit records, the job description is a particularly under-utilized resource. Most DOB applications require a licensed engineer or architect to provide a description of the scope of work when filing a permit application. This is typically a manual process using free-form text boxes or handwritten descriptions. NLP and machine learning can be used to extract semantic, topical, and other patterns from this text. Thematic structure discovery using topic modeling may complement conventional approaches (e.g., label, category, keyword) for classifying construction activities. Cities, then, could leverage this knowledge to adopt an approach similar to online product recommendation systems to provide insights into construction trends, suggest other work not considered by the applicant, and improve the efficiency of code enforcement review.

The adoption and diffusion of next-generation information technology in the public sector have begun to transform city administration (Goldsmith & Crawford, 2014; Lane, 2018; Naik, Kominers, Raskar, Glaeser, & Hidalgo, 2015). A recent survey reports that 35% of 500 U.S. cities surveyed utilize an online construction permitting system, with a 14% increase between 2015 and 2017 (William Riggs & Chavan, 2015; William Riggs & Steins, 2017). Some have suggested a cloud-based computing framework for intersectoral construction data management (Eirinaki, Dhar, & Mathur, 2016). Using NYC Open Data as an example, the researchers propose a “smart” city permit framework with a recommendation engine trained on historical permit data (Eirinaki et al., 2018). In addition, a number of studies have considered cross-domain data integration processes for improved information exchange. For example, one such study proposes a digital building permit to fill an existing gap between site-based Geographic Information System (GIS) and Building Information Modeling (BIM) data (Chognard, Dubois, Benmansour, Torri, & Domer, 2018). In addition to more efficient administration, there is significant value in applied analytics that leverage construction permit data (Hvingel, Baaner, & Schrøder, 2014). Use cases include spatial modeling for construction intensity assessment (Brandão, Correia, & Paio, 2018), spatiotemporal analysis for post-disaster recovery monitoring (Go, 2014; Stevenson, Emrich, Mitchell, & Cutter, 2010), and econometric modeling to estimate the impact of construction activity on real estate markets (Fisher, Lambie-Hanson, & Willen, 2011; Hernández-Murillo, Owyang, & Rubio, 2017; Pollakowski, 1995). Together, these studies highlight the value of construction permit data for both scientific research and real-world urban management.

At the metropolitan and regional scales, alteration of existing buildings represents an economic indicator of urban development and real estate market strength. The U.S. Census Bureau collects data on new residential construction authorized by building permits at state, metropolitan area, and county levels. Both public and private sectors, including the Department of Housing and Urban Development and the Federal Reserve Board, use construction activity as an indicator for analyzing regional economies, estimating mortgage demand, and monitoring housing investment, as well as forecasting construction industry labor markets and material demand (U.S. Census Bureau, 2018). Building alterations are a significant component of construction activities, particularly in high-density cities (Bendimerad, 2007). A near-real-time assessment of the location, extent, nature, and cost of building alterations could provide an important indicator of the health of local and regional economies beyond what could be determined by surveys alone.

In addition to the economic implications, there are increasing concerns about potential health risks associated with building alteration. In certain types of buildings, construction activities may generate health hazards due to exposure to debris and dust from lead-based paint, asbestos, or other toxins (New York State Department of Health, 2015). For example, the U.S. Environmental Protection Agency (EPA) has identified 11 target alteration actions with potential lead exposure (Battelle, 1997). Since aging building stocks are associated with both a greater likelihood of alteration activity and higher potential exposure, cities must address these public health risks, especially for vulnerable populations (Centers for Disease Control and Prevention, 1997). For example, conducted a study in NYC to investigate potential lead exposure during construction activity Reissman, Matte, Gurnitz, Kaufmann, & Leighton, (2002). The results reveal elevated blood lead levels in children living in buildings constructed before 1950 during renovation or repair work. Through more comprehensive monitoring of construction activities - including their location and scope - it would be possible to raise awareness and inform building residents and owners of potential health risks based on anticipated construction work and

Table 1
Building permit data summary.

City	Buildings[a]	Permits	Time	Frequency	Major permit types (sample size)
NYC	1,082,349	1,058,547	2000–2017	Daily	Major alteration[b] (72908) Minor alteration (816195), Minor work (228473) New construction (19463), Demolition (21508)
Los Angeles	1,140,678	573,508	2013–2017	Weekly	Alteration/ repair[b] (129051), Addition[b] (1929) HVAC[b] (63817), New (8506), Demolition (9137) Grading (11186), Plumbing (123632) Electrical (176314), Pool (1746) Fire Sprinkler (27303), Sign (8606)
Chicago	820,606	534,187	2007–2017	Daily	Renovation/alteration[b] (111825) New construction (20783), Demolition (15786) Electric wiring (193736), Easy permit (145286) Elevator equipment (13123), Sign (33648)
Austin	585,916	433,482	2000–2017	Daily	Remodel[b] (275029), Addition[b] (24578) Addition/remodel[b] (27988), Repair[b] (60375) Demolition (8010), Change out[b] (44067) Interior demolition (1617)
San Francisco	177,023	198,900	2013–2017	Weekly	Additions/ alterations/ repairs[b] (14663) New construction (349), Construction-wood (950) Over-the-counter permit (178844), Sign (3403) Demolition (600), Grade/excavate (91)
Seattle	284,017	86,051	2006–2017	Daily	Addition/ alteration[b] (59918) Tenant improvement (3141) New construction (14324) Demolition (4759), Curb cut (541)
Boston	120,994	96,977	2009–2017	Daily	Renovation-interior[b] (31475), addition[b] (1656) Change of occupancy[b] (5919), plumbing (38690) Interior/exterior work[b] (19237)

[a] Based on building footprint shapefile.
[b] Data within these permit types as major alteration.

building characteristics.

3. Data

We assemble our primary data by extracting and integrating over 2.5 million construction permit records from seven major U.S. cities: New York City (NYC), Los Angeles, Chicago, Austin, San Francisco, Seattle, and Boston. We acquire these publicly-available datasets in tabular format (.csv) from each city's open data portal as of August 2018, and exclude permits from 2018 due to the incomplete year. Table 1 summarizes the permit volume, time range, permit type, and total number of buildings for each city. The variation in permit typology reflects the lack of standardization in construction permit tracking, differing requirements in local building codes, and the nature of predominant construction-related activities in each city. While all cities have categories for new construction and demolition, alterations are ill-defined by a range of terms, including “addition”, “renovation”, “repair”, and “remodel”, as well as context-specific categories, such as “swimming pool” in Los Angeles. Seattle, as another example, is the only city that specifies “tenant improvement” as a distinct category. While specific formats and requirements may vary, cities collect many common fields, such as job type, a unique building or property identifier, a unique permit number, issuance date, estimated project cost, and a text description field. Table 2 illustrates samples from permit data in NYC. A job description summarizes the scope of work in text format, including proposed actions and major building components affected. It provides a rich bundle of information beyond the simple permit type classification.

We collect property and tax lot information as ancillary data to analyze construction activity in the context of the specific building and parcel characteristics (Table 3). Many cities maintain a land use database that contains property information with unique identifiers. In NYC, the Department of City Planning (DCP) maintains the Primary Land Use Tax Lot Output (PLUTO) database that includes lot area, Borough-Block-Lot identifier (BBL), address, building gross floor area, number of units, land use, tax assessment, and built year, among other features (NYC Department of City Planning, 2016). We merge permit data and PLUTO by BBL to identify the geo-location (latitude and longitude) of construction activity. Building identification systems vary by city; San Francisco adopted a similar system to NYC with unique block-lot numbers, while Boston uses a land parcel identifier for each tax lot. Despite these variations, ancillary data integration as a generalizable process only requires minor adjustments based on each city's building identifier system.

4. Methodology

Topic modeling is an unsupervised machine learning technique for analyzing collections of text such as news, literature, or documents (DiMaggio, Nag, & Blei, 2013; Wallach, 2006; Wang & Blei, 2011;

Table 2
Sample building alteration permit data in NYC.

BIN #	BBL	Permit type	Time	Cost ($)	Description
4159xxx	4074020xxx	Major alteration	2000/06	14,098	Legalize existing attic space as living space conjunction use with 1st fl. Legalize existing cellar toilet and partitions for home occupation.
1018xxx	1008850xxx	Major alteration	2001/01	414,779	Renovation of existing 4 story and cellar space, addition of new 3 story structure. Conversion of existing commercial to new 12 unit resident and commercial building.
1012xxx	1006427xxx	Major alteration	2001/06	34,156	Enlarge and convert existing sun room. Renovate existing kitchen and install new fixtures.

Notes: This table illustrates partial key information and does not include full permit data attributes.

Table 3
Common fields in building permit and ancillary data collected by cities.

Data	Variable name	Description
Permit data	Building ID	Unique building identifier
	Lot ID	Unique tax lot identifier
	Cost	Estimated cost
	Permit type	A categorical variable
	Issued date	A date-time variable for issuance date
	Description	Texts summarizing activities
Building & tax lot data	Building ID	Unique building identifier
	Lot ID	Unique tax lot identifier
	built area	Total gross floor area
	Floors	Number of floors
	Units	Total units
	Age	Building age by built year
	Location	Geo-location (latitude, longitude)
	Land use	Land use types or zoning class

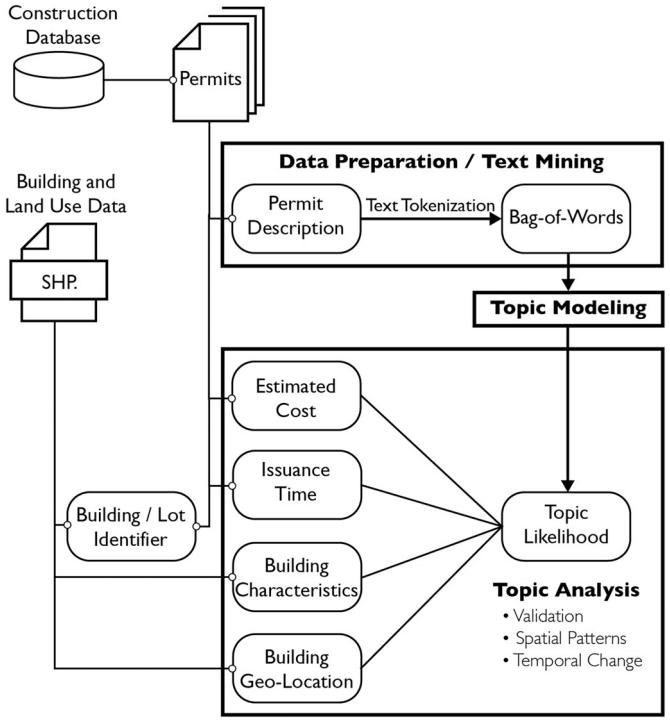

Fig. 1. Data mining and knowledge discovery framework for building construction activity.

Table 4
Frequent generic words in permit description.

Nouns	Verbs
Program, code, file, permit	Apply, issue, obtain, refer, relate
Dwell, type, job, drawing	Comply, provide, use, show, plan
Application, project, certificate	Build, submit, propose, indicate

Wang, Bowers, & Fikis, 2017). Among current approaches, latent Dirichlet allocation (LDA) is a popular topic modeling method that has been widely used in categorization and recommendation systems (Blei, Ng, & Jordan, 2003; Nguyen, Billingsley, Du, & Johnson, 2015). LDA is a probabilistic graphical model that maps each text description into a set of words and related probability of appearance based on the dictionary created from a bag-of-words (Chuang, Manning, & Heer, 2012). Previous research demonstrates that LDA can be an efficient method for thematic discovery from a large collection of text data (Hong & Davison, 2010). There are also a number of novel applications of LDA using new data sources, such as classifying social media content (Martin & Schuurman, 2017), detecting redundancy in clinical record notes based on topic similarity (Cohen, Aviram, Elhadad, & Elhadad, 2014), and extracting opinions from online reviews (Titov & McDonald, 2008). The growth of user-generated spatiotemporal data has created increasing interest in thematic patterns by place, activity, and event. Such exploratory studies include identifying the function of locations based on geo-tagged social media posts (Hu & Ester, 2013), extracting travel activity from spatial patterns of topics (Hasan & Ukkusuri, 2014), mapping the geography of public awareness (Ghosh & Guha, 2013), and tracking email communications over time (Wang & McCallum, 2006). Topic models can then be applied to news recommendations, web personalization, and social media trend detection (Jordan & Mitchell, 2015; Mobasher, 2007; Wang & Blei, 2011).

Fig. 1 illustrates our data mining and knowledge discovery process. The first step is data preparation to merge heterogeneous data sources and to clean data by removing omitted values, dropping duplicates, and correcting entry errors. Using the cleaned permit descriptions, a text mining algorithm extracts information from unstructured texts to generate structured tabulated data. The second step uses tokenized text data to build a topic model for thematic structure discovery. The final step is to join the topic modeling output with building information to discover spatial, temporal, and thematic patterns of building alteration activity. We develop this pipeline in the Python environment with multiple open-source packages.

4.1. Text mining

After data cleaning, text tokenization converts each description from a sequence of words into a list independent of grammar or syntax as a "bag-of-words" (Aldous, 1985). We use the Natural Language Toolkit ®(NLTK) to remove unnecessary words including stop words, punctuation, conjunctions, email addresses, and newline characters (Leskovec, Rajaraman, & Ullman, 2014). We also remove frequently mentioned words that do not inform alteration actions (Table 4). Based on the distribution of permit scope of work description lengths (in characters), we drop those with fewer than 23 characters (0.01 percentile). Using part-of-speech (POS) tagging, we lemmatize words to their common base form and only keep nouns (e.g., fixture, cellar, basement), verbs (e.g., move, install, renovate), and adjectives (e.g., commercial, horizontal, new). To visually inspect the relationship between words, we utilize truncated singular value decomposition (SVD) for text vectorization and dimension reduction to map words in two-dimensional space with x-y coordinates, so the distance indicates the closeness by the chance of co-occurrence in the same description (Deerwester et al., 1990).

Using the cleaned permit description data, we run a three-step text mining algorithm for information extraction to identify top words, the popularity change of words, and their co-appearance likelihood. First, a list of unique words from all descriptions is collected. Using this list, the algorithm iteratively searches through each description and appends word-appearance as key-value pairs to a dictionary. For any word that is mentioned multiple times within one description, it counts as one to prevent over-counting. We identify the most popular words by sorting total appearances in the final dictionary. Second, we run a similar counting process, but with subsets in different years to generate year-specific dictionaries of word appearances. A concatenation of all dictionaries then ranks top words per year to track popularity change over time. Finally, we use a conditional operator for each top word to generate binary variables to indicate if the description contains each top word. Correlations of these binary variables provide a statistical measure of the likelihood of word co-appearance.

4.2. Topic modeling

We adopt LDA and its terminology to organize the information

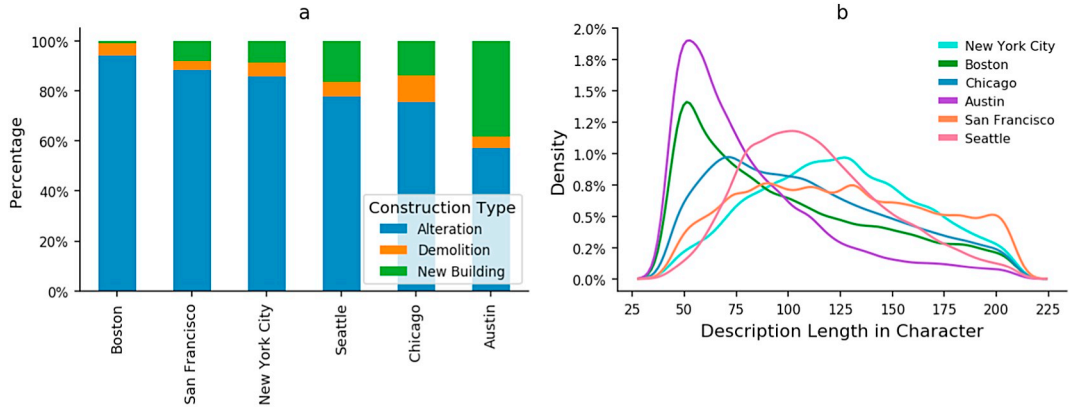

Fig. 2. (a) Citywide permits by construction type and (b) distribution of description length.

Table 5
Top salient words by topic.

	Topic 1 (renovation)	Topic 2 (addition)	Topic 3 (change of use)
NYC	**Plumbing** partition **interior** install **fixture** **renovation** convert minor **kitchen** unit **bathroom** change cellar room work plan **mechanical** stair floor wall roof **renovate** remove basement **boiler** finish construction exterior window **plumb**	Conversion propose **extension** rear **addition** enlargement **vertical horizontal** family residence building occupancy legalize **add** cellar plan attic legalization build floor renovation **construct** submit frame **erect** yard **deck bedroom** house **porch**	**Change convert** building **office store** file cellar amend **residential** floor **legalize** obtain basement **apartment commercial** create construction **occupancy** build extend single remove **eat** apt bulk **medical** accessory conjunction road **retail**
Boston	Kitchen floor **bathroom** wall paint **renovate** tile **cabinet electrical** unit ceiling remove **interior** remodel **plumbing bath** door room finish partition drywall demo plaster **light fixture** carpet vanity add **hvac shower** sprinkler **retrofit** soft change ordinance maher **compliance seismic upgrade** require wall revision	Window repair **structural** door change **remove** **exterior roof basement** rear **deck** wall **porch** damage **stair** vinyl trim water wood entry house interior need building board paint insulation **frame concrete rebuild addition** rear bathroom **deck remodel bedroom roof** window **stair** interior kitchen room	**Office** include sprinkler **change** interior renovate **occupancy** permit **restaurant egress** unit alarm room **store retail** finish minor build **tenant** equipment upgrade apartment demolition mechanical single addition **residential** scope lobby **storage** unit **ordinance** addition **legalize** kitchen antenna ground remove **legalization residential** ord nov
San Francisco	Separate **mandatory** alarm construction include building structural door exterior provide office work retail finish service restaurant floor nfpa	Level garage road **horizontal floor** door bath **exterior** renovate **basement** yard **vertical remove** **facade** rehab ground finish wall	Propose **convert** adu comply **storage** floor cabinet studio **equipment** accessory building roof add modify wall maher build panel
Chicago	Wall **electrical** room revision antenna equipment remove door site **kitchen bathroom** window addition **repair** exterior change roof previous **plumbing mechanical** unit facility associate **fixture** service structural masonry **upgrade** ceiling parking	Single **basement addition** interior residence rear **deck frame** residential **stair** eration unit **garage** wood masonry brick **erect roof** open repair renovation plumbing electrical **exterior** new **remodel** violation car building dormer	Interior eration **office** self inspection conditional cert **tenant** correction plumbing renovation electrical floor **change** build **retail** audit **suite** partition **restaurant** mechanical **business occupancy** certified **commercial** propose certify demolition work include
Seattle	Close incomplete expire accessory garage addition alter **seismic** remove basement wall **voluntary** parking family structure rear detach detached foundation **retrofit** convert **upgrade** single retain complete **inspection** final portion review expired	**Addition** interior **deck basement** inspection repair window kitchen bathroom **roof remodel** **exterior** main **structural remove** door **stair** wall **porch** replace non damage unit floor replacement **bedroom** kind enclose entry portion	**Tenant** improvement **office** occupy **commercial** change interior **retail restaurant** portion structural minor non **suite** antenna **facility** corner communication service center **equipment** road **medical rooftop** mechanical **warehouse mix** utility include floor
Austin	**Elec** residence service **hvac** new **water** exist residential **heater upgrade** eud line **repair** service **watt** story **gas** loop meter city **plumbing** refer home **electrical** outside sf **replace** austin commercial	**Addition remodel** bathroom kitchen **add** garage room **porch** window residence wall **bedroom** interior sf repair **partial exterior** door demo **deck** **remove** plumbing **rear roof** create closet electrical cover light wood	Remodel interior **office change tenant** create **retail** finish duct work complete admnbus air heat central service **restaurant** prof finish **suite medical** personal unit **residential** apt administrative **sale** restroom **apartment** building

Words in bold are relatively distinct in a topic group and consistent across different cities.

hierarchy as words, document, corpus, and topics (Blei et al., 2003). A topic is an abstract subject according to the pattern of bags of words in documents of a corpus. The topics are "latent" since there is an unknown number of topics (Blei, 2012). In LDA, each topic represents a probability distribution of words collected from the corpus, and a document represents a probability distribution of topics (Steyvers & Griffiths, 2007). LDA summarizes each topic as a *N*-dimension vector with different probability-of-appearance for all unique words. Since the identified topics are not explicit, LDA relies on salient words with the highest probabilities to represent a semantic theme for each topic group. These salient words are not mutually exclusive to a single topic. The LDA model output quantifies the thematic composition of each description with *K* different percentages that sum up to 1, representing the proportions of all *K* topics in a description.

We use pyLDAvis® to inspect model outputs and identified topics (Mabey, 2018; Rehurek, 2018; Rehurek & Sojka, 2011; Sievert & Shirley, 2014). The LDA model objectively quantifies the thematic composition of each description based on its topic likelihood. The results describe each description as an array of probabilistic measures summarizing the proportion of different themes (Griffiths & Steyvers, 2004). If the thematic composition is not even, the primary theme is the one with the highest probability. For example, if a permit contains 14% topic 1, 59% topic 2, and 27% topic 3, we infer it is predominantly related to topic 2. LDA is appropriate in this analysis since each alteration project may involve multiple, complementary activities. LDA requires a predefined number of latent topic *k*, which is usually decided based on domain knowledge, research objectives, or practical use cases (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). We choose three

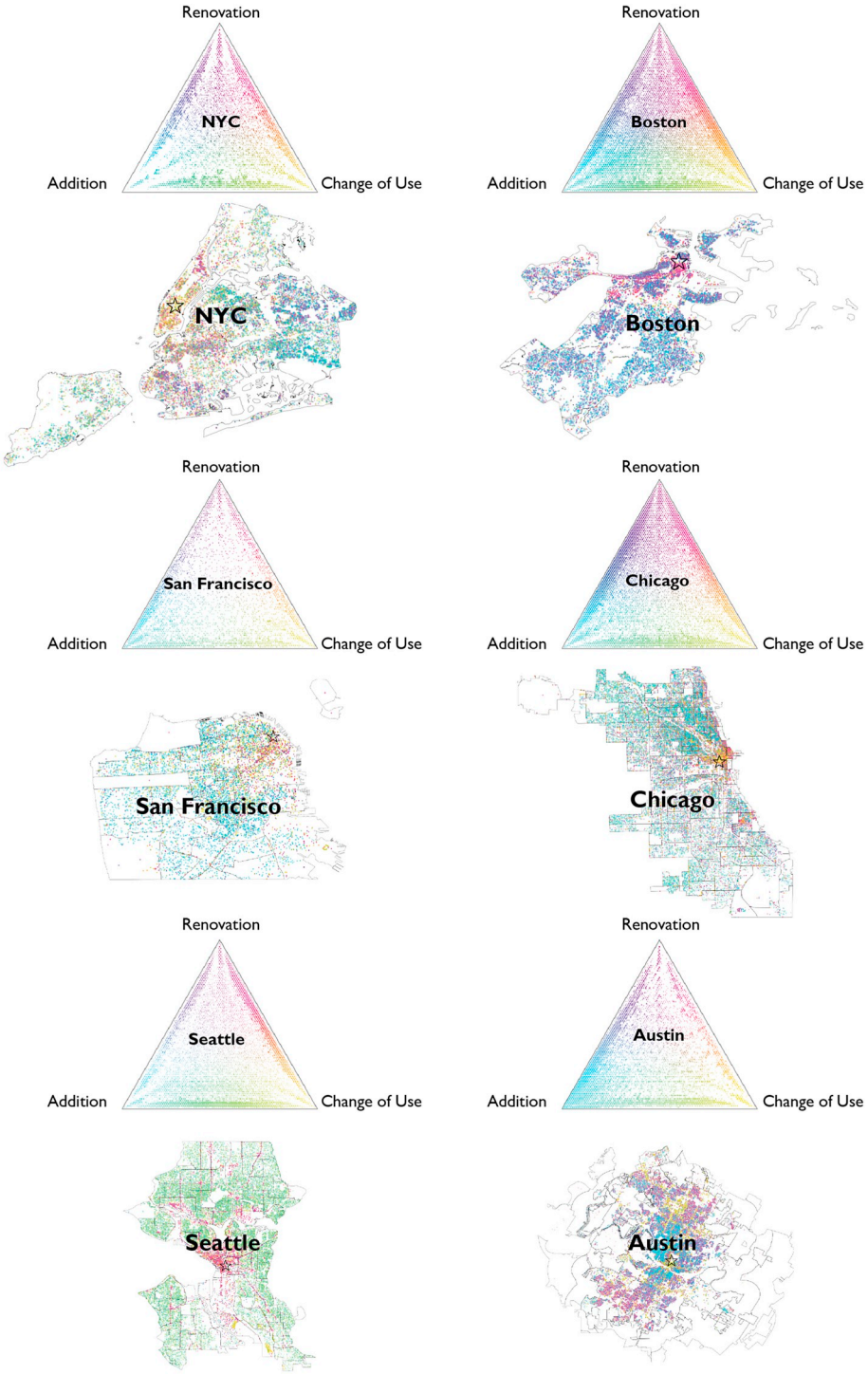

Fig. 3. Simplex plots and maps visualize thematic patterns of building alterations.

topics to align with major alteration activities related to renovation, change of use, and addition. A common practice is to start with a relatively smaller number of topics and then to inspect how salient words appear across topics. To identify different themes, we summarize each topic with salient words by ranking frequency of appearance. These top salient words indicate a theme for each topic (Zhao et al., 2015). For example, if a topic has top words “renovate”, “retrofit”, and “upgrade”, its theme is more likely related to renovation.

Since a topic model quantifies each description with k probabilistic measures as compositional data, we use Plotly® to generate ternary plots based on each permit's topic probabilities as a visualization of its thematic composition (Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015). A ternary plot graphically maps each data point in an equilateral triangle according to its three variables v_1, v_2, v_3 (Howarth, 1996). This method can extend for k topics by visualizing data into a k-1 simplex geometry with k vertices representing all topics. We rescale topic probabilities π_{i1}, π_{i2}, and π_{i3} into a range between 0 and 255, such that three probabilities can be mapped as a single value in an RGB (Red-Green-Blue) color scheme.

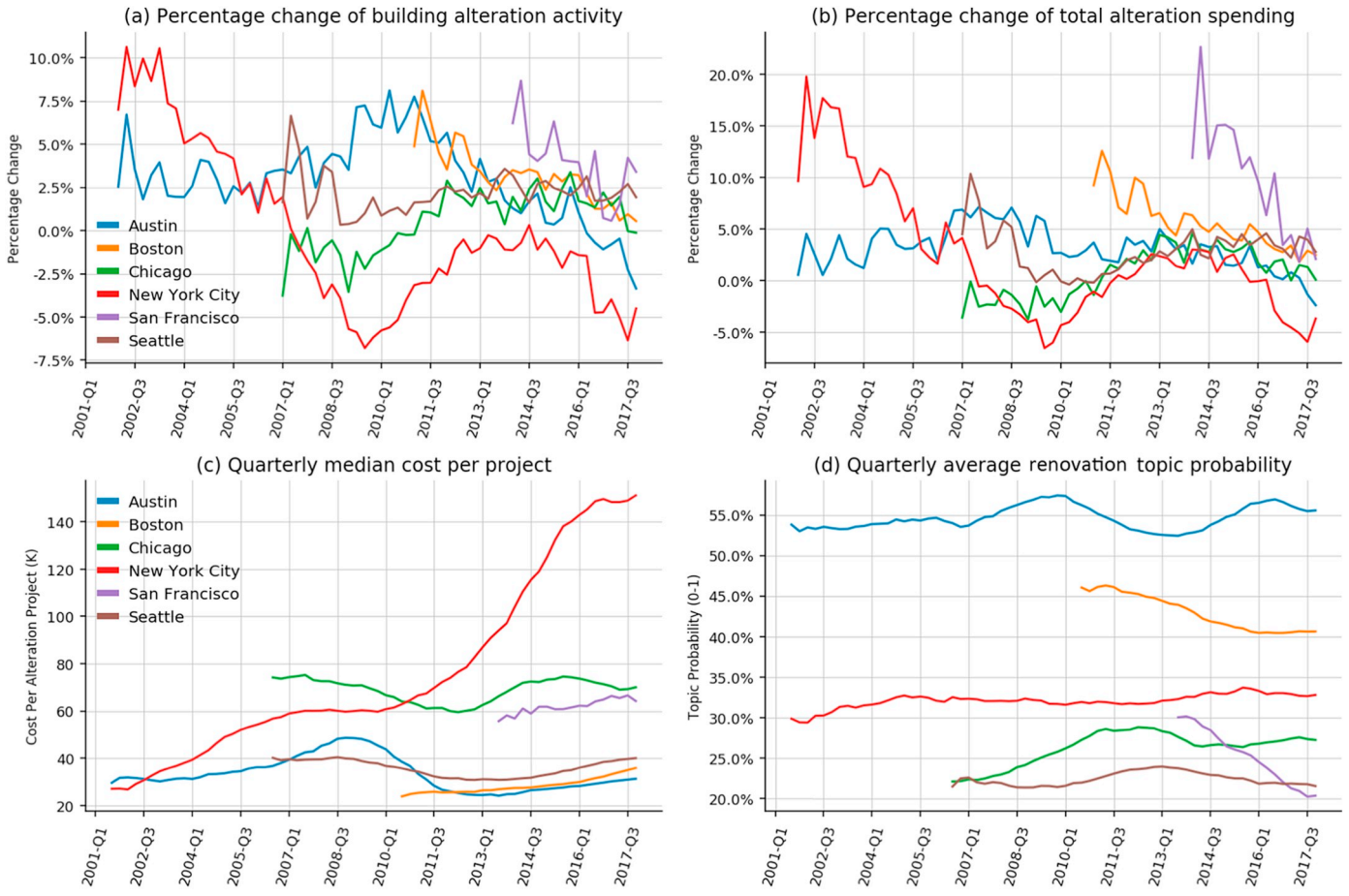

Fig. 4. Time-series of (a) alteration growth, (b) total alteration spending growth, (c) monthly median cost per project in six cities, and (d) monthly average renovation topic probability.

5. Results

5.1. Construction permit database

We build a comprehensive database of construction projects (n = 1,408,339) with common features, including permit identifier, issuance date, permit type, estimated cost, latitude and longitude, as well as the city-specific building identifier and parcel identifier. We preserve the original permit typology defined in Table 1 and regroup data into three categories as alteration, new construction, and demolition. A cross-city comparison shows the relative proportions of the three construction types (Fig. 2a). In Boston, alterations account for more than 90% of total building permits issued. In contrast, Austin has roughly equal percentages of new construction and renovation (40% of total permits, respectively), indicating how these data can be used to understand localized patterns of urban development and redevelopment. The text length of permit descriptions ranges from 20 to 250 characters across the studied cities (Fig. 2b). Los Angeles only allows 75 characters as a maximum length, possibly due to its online filing system settings; therefore, we do not include Los Angeles for topic modeling. After removing the unnecessary space, conjunctions, and generic terms defined in Table 4, we collect 54,023 unique words with 4,448,998 appearances in total. Top salient words include terms describing which building components are involved (e.g. plumbing, fixture, frame, water), where alteration work occurs (e.g. kitchen, bedroom, bathroom, porch, garage), and what specific actions are proposed (e.g. convert, renovate, legalize, replace, remodel, repair).

5.2. Alteration thematic discovery

Since LDA does not explicitly identify topics, top salient words for each group indicate the underlying theme (Table 5). LDA models do not assign each word exclusively to a single topic, so a word may appear in more than one topic group. Assuming a topic model captures context with salient words, it is necessary to interpret each word along with other words within the same topic group. Some words (in bold) are relatively distinct in a topic group and consistent across different cities, indicating three unique thematic groups of alterations. Topic 1 relates to *renovation*, which can include upgrading building systems, replacing fixtures, or adding new equipment; Topic 2 relates to *addition* that typically involves changing the physical structure or bulk of the current building, such as adding floors or expanding existing floor area; and Topic 3 relates to *change of use*, such as converting office space to residential or adding a new food establishment use. We can describe an alteration activity by the probabilistic distributions of these three major alteration themes. This measure may better represent alteration activities in practice, since a project will often involve some aspects of renovation, addition, or change of use in varying proportions.

The visualizations below reveal spatial patterns of alteration activities and the influence of urban form (density, distance to urban center, or street network), building typology (single-family, multifamily, or condominiums), zoning regulations, and ownership types (Fig. 3). Since not every city defines its central business district (CBD) with explicit boundaries, we mark landmarks that commonly define the city center (indicated as stars in Fig. 3). These landmark buildings are: the Empire State Building (NYC), Government Center (Boston), Transamerica Pyramid (San Francisco), Willis Tower (Chicago), Columbia Center

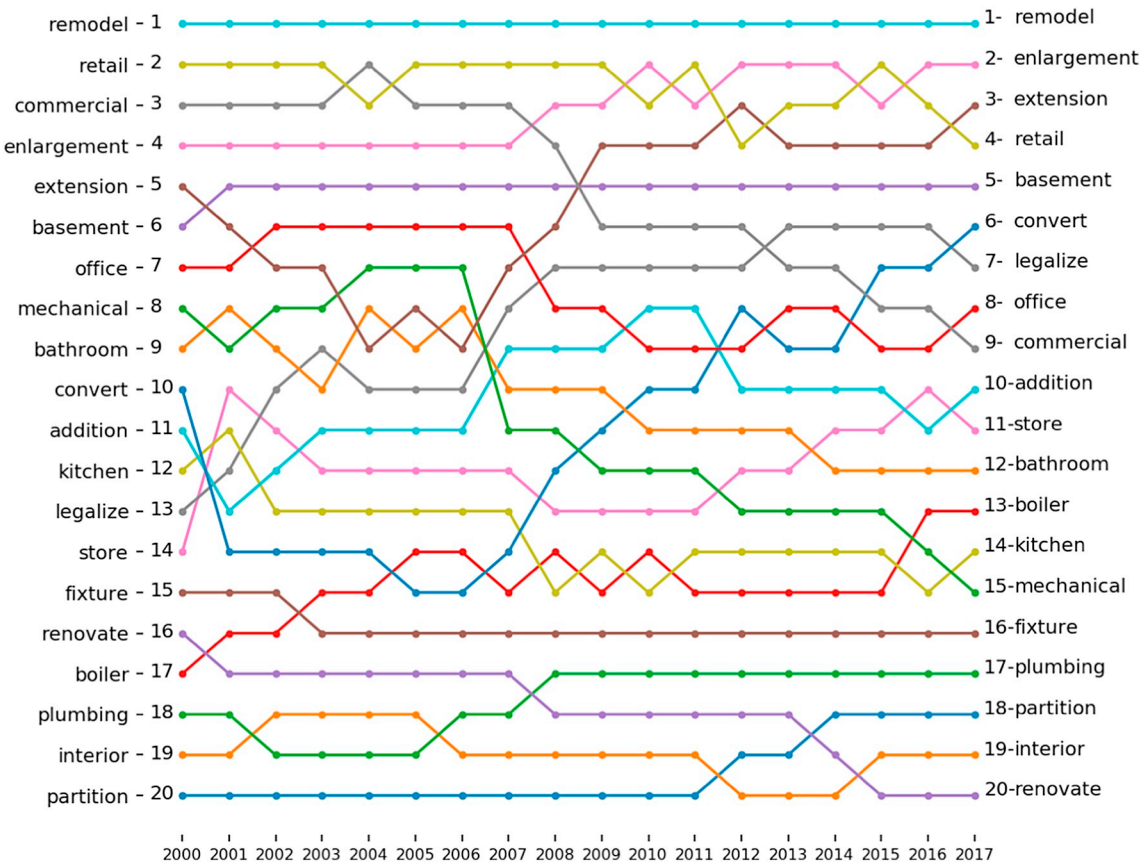

Fig. 5. Annual topic salient words ranking by appearance, New York City.

(Seattle), and Texas Capitol (Austin). In Boston, Chicago, and Seattle, there is a clear thematic pattern associated with the distance to the CBD. This pattern indicates more renovation-dominated activity in the urban center, possibly due to the older, existing building stock and higher property values. Given the multitude of factors influencing the location and extent of alteration activity, the spatial representations are intended to provide only illustrations of the geographic distribution of thematic groups.

Time-series analysis reveals the long-term changes in construction activities in different cities. Fig. 4a shows the quarterly percentage change of building alteration permits issued. Austin, Seattle, Boston, and San Francisco experienced a positive change over time, indicating sustained growth of alteration activities. The alteration growth rate in Austin peaked in 2010 and then decreased, following a similar trend as in Boston and San Francisco. NYC, however, has a negative change rate since 2007 that implies decreasing alteration activity over time. This trend is possibly due to the impact of the subprime mortgage crisis and resulting economic recession beginning at the end of 2007. Fig. 4b shows the percentage change of total alteration spending as a proxy for overall alteration intensity, accounting for both permit volume and estimated cost. Contrary to Fig. 4a, total spending in NYC returned to positive growth from 2012 to 2016, possibly due to the increased average construction cost per project, while remaining relatively constant or declining in the other cities. Fig. 4c measures median spending per project as a proxy for construction costs in different cities. An increasing cost-per-project can be the result of higher material costs, rising labor costs, longer project duration, and larger scope of work. While a majority of cities experienced relatively constant per project costs, NYC witnessed an approximately five-fold increase between 2001 and 2017. Fig. 4d shows the monthly average renovation probability per project, which is the average likelihood of renovation compared to all other alteration types. The overall average topic probabilities indicate the prevalence of renovation actions involved in building alterations (Austin > Boston > NYC > Chicago > San Francisco > Seattle) according to topic modeling output. The results indicate that retrofitting probability keeps stable with seasonal fluctuations in most cities, while San Francisco and Boston demonstrate a decreasing trend in renovation probability per project.

The text mining process computes the annual total appearances per word to rank topic salient words as an indicator of topic changes over time. Using NYC as an example, Fig. 5 shows the prevalence of certain words changed dramatically from 2000 to 2017. While “remodel” is stable as the most popular term, “extension”, “convert”, and “legalize” have gained popularity as top salient words. There is also an increasing incidence of “boiler”, perhaps reflecting the impact of a recent local law requiring the replacement of older, heavy fuel oil boilers.

5.3. Validation

Topic models generate probabilistic output that estimates the likelihood of alteration decisions. This provides insights for cities, like NYC, where the permit typology is not defined by alteration actions. While NYC does not define permit types as “renovation” or “addition”, it does provide predefined check-boxes for applicants to indicate specific actions (e.g., boiler, plumbing, horizontal enlargement) involved in the proposed alteration. Our model results reveal how a specific decision is associated with different topic likelihood distributions (Fig. 6). The results show that alterations associated with mechanical and boilers have a higher topic likelihood of renovation. In contrast, alterations with enlargement have a higher topic likelihood of addition. In fact, horizontal enlargement and vertical enlargement are not commonly found in the same permit application, evidenced by the moderate correlation (corr. = 0.3) between the respective check-box binary variables. The similar topic likelihood distributions indicate equivalent

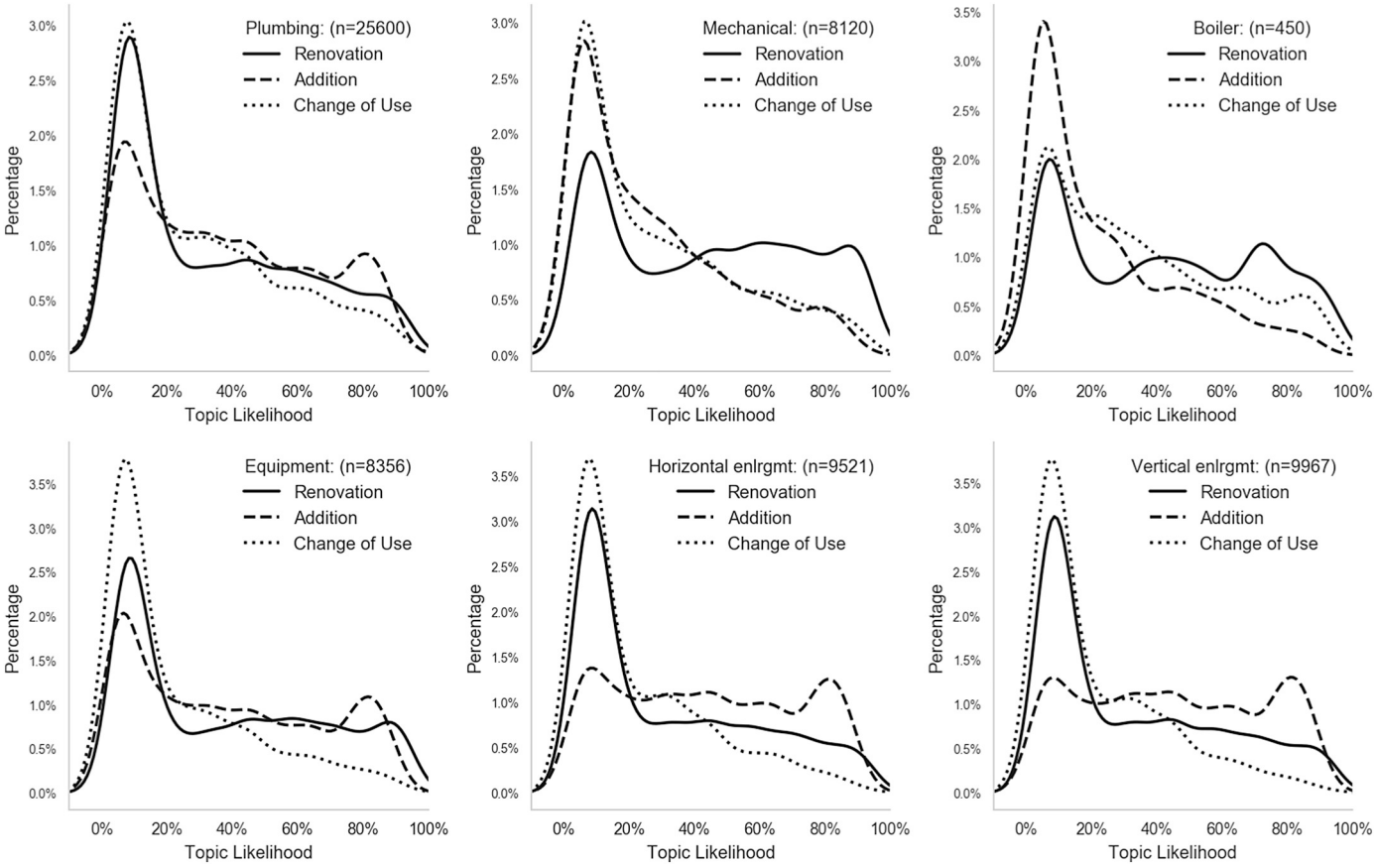

Fig. 6. Topic likelihood distributions by alteration decisions and check-box selection, New York City.

Table 6
Building alteration topic composition in Boston and Austin.

City	Average topic likelihood Original category	Renovation	Addition	Change of use
Boston	Renovations-interior	56.3%	26.0%	17.7%
	Addition	16.7%	37.8%	45.5%
	Interior/exterior work	25.0%	56.1%	18.9%
	Change occupancy	13.2%	15.6%	71.2%
Austin	Repair	53.9%	36.3%	9.8%
	Change out	60.2%	8.0%	31.8%
	Remodel	29.0%	23.8%	47.3%
	Addition	11.1%	80.8%	8.1%
	Addition and remodel	7.5%	84.7%	7.8%

likelihoods of renovation, addition, and change of use for projects with either horizontal or vertical enlargement.

For cities where permit categories are based on specific actions, the descriptive summary of topic likelihood per permit category provides a validation of topic modeling output. Table 6 demonstrates different topic compositions based on the originally defined alteration categories in Boston and Austin. Results reveal a relationship between the original category selection and the dominating topic (e.g. category "Renovations-Interior" contains 56.3% textual content related to renovation). The topic modeling outputs can be used to identify construction activities that initially selected other categories, but contain a high likelihood of renovation work.

6. Discussion and implications

Topic modeling provides new insights by quantifying the composition of construction activity based on the thematic structure of permit descriptions. Conventional categorization has an explicit limitation since it allocates each permit to a single group. In reality, these project types are not mutually exclusive. For instance addition (changing the structure of the building by adding rooms, removing walls, or expanding the square footage) and renovation (upgrading building systems and home improvement) work may occur concurrently in a single project. For such reasons, few cities provide clear definitions for each building alteration category. Table 7 summarizes the terminology used to describe building alteration activity and how these vary by source. This analysis provides two key insights. First, currently there is no universal definition for building alteration, and its explanation varies with commercial, administrative, and legal context. Second, it may be impractical to create a rigid classification system for alteration activity, since alteration projects often involve multiple intersecting and interrelated actions. Therefore, a topic model has the unique benefit of quantifying each permit description based on its composition of underlying topics, which enables a dynamic categorization for pattern discovery.

Admittedly, this study has several limitations to be addressed in future work. First, since LDA is a unsupervised generative probabilistic approach, the number of latent topics k needs to be predefined. Although a statistically-best k derives from optimizing topic coherence values, this approach often generates an excessive number of topics that are too complex in practice. On the other hand, a smaller k may lead to a model that is too coarse to identify distinct topics (Zhao et al., 2015). Since this study aims to objectively decompose each description by a fixed number of thematic groups, we argue that topic distinctiveness is relative and a smaller k yields an appropriate balance between practical application and thematic complexity for output interpretation and visualization. Second, LDA topic modeling maps each description with k topic probabilities, but it does not capture correlation among topics. To address the above limitations, it is possible to optimize k based on a

Table 7
Building alteration terminology defined by various sources.

Terminology	Definition	Source
Renovation	Renovation may involve remodeling, renewal of an outdated or damaged structure or associated equipment and materials, or partial demolition and any reconfiguration or replacement of interior partitions.	NYC DOB
	A renovation means you're updating an existing structure with cosmetic changes.	Realtor.com
	Renovating means to make new again. An out-of-date kitchen, updated with new finishes and fixtures, has been renovated. Replacing old windows with new ones is a renovation project.	Zillow.com
	This can include everything from restoring an older home to a dated kitchen renovation or garage remodel to a dreary bathroom.	homeadvisor.com
	Extensive alteration work in addition to work on the exterior shell of the building and/or primary structural components and/or the core and peripheral MEP and service systems and/or site work.	U.S. Green Building Council
Addition & remodel	An extension or modification to an existing house, which may include a second-story addition, dormer, footprint expansion, interior reconfiguration, or house lift.	Seattle Dept. of Construction & Inspections
Addition	This includes adding new rooms or a sunroom to a home, but it also encompasses the construction of patios, porches and decks as well, depending on local codes. Enclosing a garage may be considered an addition because it would increase the heated space of the home.	homeadvisor.com
Remodeling	A remodel involves changing the structure through demolition and construction.	Realtor.com
	Broadly describes any kind of change to an existing house, including change the character of a house or a portion of a house.	Zillow.com
	An interior remodel fixes or updates the appearance or use of an existing home without adding or removing square footage.	City of Austin
Structural changes	Structural changes generally involve alterations to the bones of the structure, including the addition or removal of walls or finishing an attic or basement space.	homeadvisor.com
Conversion	Converting an attached porch, carport or garage into interior living space.	City of Austin
Change use	If your new business is different than the former business that was in the space for example, you want to open a retail store in a space that was previously an office.	Seattle Dept. of Construction & Inspections

topic tree with hierarchical relationships using Bayesian non-parametric models (Blei, 2012). Third, this study assumes that text descriptions are independent of time and location, which is a common assumption in topic modeling. However, recent studies, and our own results, indicate that topics may shift along time, which ultimately requires a "Topic-over-Time" (ToT) model that integrates both text and time into a generative process for more precise topic classification (Dubey, Hefny, Williamson, & Xing, 2013). Finally, permit job descriptions are relatively short compared to other text data, such as news articles, email messages, or customer review comments. Short descriptions create sparsity in the documents, which can constrain the performance of topic modeling. In this study, we deliberately drop the case of Los Angeles due to its very short permit description (with an average length of 60 characters compared to 120 in other cities). Interestingly, Twitter data shares similar characteristics in that it has limited text length (no more 140 characters before 2017 and now expanded to 280) (Naveed, Gottron, Kunegis, & Alhadi, 2011). Recent studies investigate various techniques to improve topic models for short text data, by training a model on aggregated documents (Hong & Davison, 2010) or clustering distributed representations of words (Sridhar, 2015). Therefore, future work will explore various NLP techniques to improve topic modeling for sparse documents.

Since identifying the theme for each topic relies on subjective human interpretations, it requires domain knowledge from the building and construction industry. Previous research indicates that domain knowledge related to the likelihoods of word co-appearance can be used as a prior to improve LDA models (Andrzejewski, Zhu, & Craven, 2009). Therefore, we plan to invite a group of experts, including DOB officials, architects, and contractors to further validate the topic model and improve our knowledge discovery framework.

In practice, this study contributes to data-driven methods for cities to identify the scope and spatiotemporal patterns of alteration activity from building permit records. As alteration has surpassed new construction as the dominant development activity in most developed countries, cities need better insight and understanding of alteration activities, particularly considering the impact of existing buildings on global carbon emissions, public health, and economic growth (Bloomberg & Pope, 2017; Kontokosta, 2016; Sartori et al., 2008). Our results demonstrate the potential of using NLP to develop a generalizable and flexible algorithm to measure alteration decisions. This work provides the methodological foundations for near-real-time construction activity analytics. Our model is built upon extensive data processing and exploratory analysis to identify common variables across all seven studied cities. Although this analytical pipeline creates a generalizable process, we do not suggest that machine learning should replace current permit categorization. Instead, we provide a complementary approach to classify alteration activity while preserving existing typologies and recognizing the implications of each city's legal regularly, administrative, and technical context.

7. Conclusion

Our exploratory analysis attempts to identify the thematic structure of building alteration activities using construction permit data and scope of work descriptions. We develop a generalizable and reproducible approach to extract common variables and process text data from digital permit records in multiple cities. Ultimately, this study contributes to the growing literature on urban data science and artificial intelligence in city management. By introducing a novel application of NLP to an under-utilized urban data source, we aim to support data-driven decision-making through high-resolution modeling of the dynamics of the built environment.

This study demonstrates the potential of latent topic modeling for building alteration activities based on permit descriptions. Topic modeling enables a probabilistic measure of complex building alteration decisions, which can enable future recommendation and analyticals systems. We introduce a generalizable analytical pipeline to seven major cities in the U.S. to test the feasibility and value of analyzing building alteration permit data. The results show the power of data mining and modeling using text-rich alteration descriptions along with ancillary building and land parcel information. This approach may assist cities to better monitor building alteration activity, analyze spatiotemporal patterns, and more fully understand the economic, social, and environmental implications of changes to the urban built environment.

Acknowledgments

This material is based upon work supported by the National Science Foundation, Grant #1653772.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2019.101383.

References

Aldous, D. J. (1985). Exchangeability and related topics. *École d' Été de Probabilités de Saint-Flour XIII1983* (pp. 1–198). Springer. https://doi.org/10.1007/BFb0099421.

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). ACM. https://www:ncbi:nlm:nih:gov/pmc/articles/PMC2943854/.

Battelle (1997). Lead exposure associated with renovation and remodeling activities: Environmental field sampling study. Technical Report U.S. Environmental Protection Agency https://www:epa:gov/sites/production/files/documents/r96-007:pdf.

Beccali, M., Cellura, M., Fontana, M., Longo, S., & Mistretta, M. (2013). Energy retrofit of a single-family house: Life cycle net energy saving and environmental benefits. *Renewable and Sustainable Energy Reviews, 27*, 283–293. https://doi.org/10.1016/j.rser.2013.05.040.

Bendimerad, A. (2007). *Developing a leading indicator for the remodeling industry*. Joint Center for Housing Studies of Harvard Universityhttps://www:jchs:harvard:edu/sites/default/files/n07-1 bendimerad 0:pdf.

Bettencourt, L. M. (2014). The uses of big data in cities. *Big Data, 2*, 12–22. https://doi.org/10.1089/big.2013.0042.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://dl:acm:org/citation:cfm?id=944919:944937.

Bloomberg, M., & Pope, C. (2017). *Climate of hope: How cities, businesses, and citizens can save the planet.* St. Martin's Press.

Brandão, F., Correia, R., & Paio, A. (2018). Measuring urban renewal: A dual kernel density estimation to assess the intensity of building renovationcase study in Lisbon. *Urban Science, 2*, 91. https://doi.org/10.3390/urbansci2030091.

Centers for Disease Control and Prevention (1997). Children with elevated blood lead levels attributed to home renovation and remodeling activities–New York, 1993–1994. *Morbidity and Mortality Weekly Report, 45*, 1120. https://www:cdc:gov/mmwr/preview/mmwrhtml/00045033:htm.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems* (pp. 288–296). . http://papers:nips:cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models:pdf.

Chognard, S., Dubois, A., Benmansour, Y., Torri, E., & Domer, B. (2018). Digital construction permit: A round trip between GIS and IFC. In I. F. C. Smith, & B. Domer (Eds.). *Advanced computing strategies for engineering* (pp. 287–306). Cham: Springer International Publishing. https://link:springer:com/chapter/10:1007/978-3-319-91638-5 16.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the international working conference on advanced visual interfaces* (pp. 74–77). ACM. https://doi.org/10.1145/2254556.2254572.

Cohen, R., Aviram, I., Elhadad, M., & Elhadad, N. (2014). Redundancy-aware topic modeling for patient record notes. *PLoS One, 9*, e87555. https://doi.org/10.1371/journal.pone.0087555.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9.

De Melo, A. B., Goncalves, A. F., & Martins, I. M. (2011). Construction and demolition waste generation and management in Lisbon (Portugal). *Resources, Conservation and Recycling, 55*, 1252–1264. https://doi.org/10.1016/j.resconrec.2011.06.010.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics, 41*, 570–606. https://doi.org/10.1016/j.poetic.2013.08.004.

Dubey, A., Hefny, A., Williamson, S., & Xing, E. P. (2013). A nonparametric mixture model for topic modeling over time. *Proceedings of the 2013 SIAM international conference on data mining* (pp. 530–538). SIAM. https://doi.org/10.1137/1.9781611972832.59.

Eirinaki, M., Dhar, S., & Mathur, S. (2016). A cloud-based framework for smart permit system for buildings. *Proceedings of the 2016 IEEE international smart cities conference (ISC2)* (pp. 1–4). . https://doi.org/10.1109/ISC2.2016.7580821.

Eirinaki, M., Dhar, S., Mathur, S., Kaley, A., Patel, A., Joshi, A., & Shah, D. (2018). A building permit system for smart cities: A cloud-based framework. *Computers, Environment and Urban Systems, 70*, 175–188. https://doi.org/10.1016/j.compenvurbsys.2018.03.006.

Fayyad, U., & Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Communications of the ACM, 45*, 28–31.

Fisher, L., Lambie-Hanson, L., & Willen, P. S. (2011). A profile of the mortgage crisis in a low-and-moderate-income community. *The American Mortgage System: Crisis and Reform,* 137–158.

Ghosh, D., & Guha, R. (2013). What are we `tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science, 40*, 90–102. https://doi.org/10.1080/15230406.2013.776210.

Go, M. H. (2014). The power of participation: Explaining the issuance of building permits in post-Katrina New Orleans. *Urban Affairs Review, 50*, 34–62. https://doi.org/10.1177/1078087413476462.

Goldsmith, S., & Crawford, S. (2014). *The responsive city: Engaging communities through data-smart governance.* John Wiley & Sons.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Harford, T. (2014). Big data: A big mistake? *Significance, 11*, 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x.

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies, 44*, 363–381. https://doi.org/10.1016/j.trc.2014.04.003.

Helms, A. C. (2003). Understanding gentrification: An empirical analysis of the determinants of urban housing renovation. *Journal of Urban Economics, 54*, 474–498. https://doi.org/10.1016/S0094-1190(03)00081-0.

Hernández-Murillo, R., Owyang, M. T., & Rubio, M. (2017). Clustered housing cycles. *Regional Science and Urban Economics, 66*, 185–197. https://doi.org/10.1016/j.regsciurbeco.2017.06.003.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. *Proceedings of the first workshop on social media analytics* (pp. 80–88). ACM. https://doi.org/10.1145/1964858.1964870.

Howarth, R. J. (1996). Sources for a history of the ternary diagram. *The British Journal for the History of Science, 29*(3), 337–356. https://doi.org/10.1017/S000708740003449X.

Hu, B., & Ester, M. (2013). Spatial topic modeling in online social media for location recommendation. *Proceedings of the 7th ACM conference on recommender systems* (pp. 25–32). ACM. https://doi.org/10.1145/2507157.2507174.

Hvingel, L., Baaner, L., & Schrøder, L. (2014). Mature e-government based on spatial data-legal implications. *International Journal of Spatial Data Infrastructures Research, 9*, 131–149. https://doi.org/10.2902/1725-0463.2014.09.art6.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255–260. https://doi.org/10.1126/science.aaa8415.

Juan, Y.-K., Gao, P., & Wang, J. (2010). A hybrid decision support system for sustainable office building renovation and energy performance improvement. *Energy and Buildings, 42*, 290–297. https://doi.org/10.1016/j.enbuild.2009.09.006.

Kontokosta, C. E. (2013). Energy disclosure, market behavior, and the building data ecosystem. *Annals of the New York Academy of Sciences, 1295*, 34–43. https://doi.org/10.1111/nyas.12163.

Kontokosta, C. E. (2016). Modeling the energy retrofit decision in commercial office buildings. *Energy and Buildings, 131*, 1–20. https://doi.org/10.1016/j.enbuild.2016.08.062.

Kontokosta, C. E. (2018). Urban informatics in the science and practice of planning. *Journal of Planning Education and Research*. https://doi.org/10.1177/0739456X18793716.

Ku, M., & Gil-Garcia, J. R. (2018). Ready for data analytics? Data collection and creation in local governments. *Proceedings of the 19th annual international conference on digital government research: Governance in the data age* (pp. 36). ACM. https://doi.org/10.1145/3209281.3209381.

Lane, J. (2018). Building an infrastructure to support the use of government administrative data for program performance and social science research. *The Annals of the American Academy of Political and Social Science, 675*, 240–252. https://doi.org/10.1177/0002716217746652.

Lees, L. (2003). Super-gentrification: The case of Brooklyn Heights, NEW YORK city. *Urban Studies, 40*, 2487–2509. https://doi.org/10.1080/0042098032000136174.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets.* Cambridge University Presshttps://dl:acm:org/citation:cfm?id=2124405.

Mabey, B. (2018). pyldavis. https://pyldavis:readthedocs:io/en/latest/readme:html.

Martin, M. E., & Schuurman, N. (2017). Area-based topic modeling and visualization of social media for qualitative GIS. *Annals of the American Association of Geographers, 107*, 1028–1039. https://doi.org/10.1080/24694452.2017.1293499.

Mobasher, B. (2007). Data mining for web personalization. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.). *The adaptive web: Methods and strategies of web personalization* (pp. 90–135). Springer. https://doi.org/10.1007/978-3-540-72079-93https://doi:org/10:1007/978-3-540-72079-93.

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2015). *Do people shape cities, or do cities shape people? The co-evolution of physical, social, and economic change in five major us cities.* National Bureau of Economic Researchhttps://doi.org/10.3386/w21620.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching microblogs: Coping with sparsity and document quality. *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 183–188). ACM.

Neef, D. (2014). *Digital exhaust: What everyone should know about big data, digitization and digitally driven innovation.* Pearson Education.

New York State Department of Health (2015). Lead exposure during renovation and remodeling. https://www:health:ny:gov/environmental/lead/renovations and remodeling:htm.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, 3*, 299–313. https://doi.org/10.1162/tacl.a.00140.

NYC Department of City Planning (2016). Pluto and mappluto. https://www1:nyc:gov/site/planning/data-maps/open-data/dwn-pluto-mappluto:page.

Offenhuber, D., & Ratti, C. (2014). *Decoding the city: Urbanism in the age of big data.* Birkhäuser.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and*

analysis of compositional data. John Wiley & Sons.
Pollakowski, H. O. (1995). Data sources for measuring house price changes. *Journal of Housing Research, 6*, 377–387. http://www:jstor:org/stable/24832837.
Rehurek, R. (2018). Gensim: Topic modeling for humans. https://radimrehurek:com/gensim/.
Rehurek, R., & Sojka, P. (2011). Gensimstatistical semantics in python. https://www:fi:muni:cz/usr/sojka/posters/rehurek-sojka-scipy2011:pdf.
Reissman, D. B., Matte, T. D., Gurnitz, K. L., Kaufmann, R. B., & Leighton, J. (2002). Is home renovation or repair a risk factor for exposure to lead among children residing in New York city? *Journal of Urban Health, 79*, 502–511. https://doi.org/10.1093/jurban/79.4.502.
Sartori, I., Bergsdal, H., Müller, D. B., & Brattebø, H. (2008). Towards modelling of construction, renovation and demolition activities: Norway's dwelling stock, 1900–2100. *Building Research & Information, 36*, 412–425. https://doi.org/10.1080/09613210802184312.
Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., et al. (2012). Linked open government data: Lessons from data. gov. uk. *IEEE Intelligent Systems, 27*, 16–24. https://eprints:soton:ac:uk/340564/.
Sievert, C., & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). .
Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 192–200). .
Stevenson, J. R., Emrich, C. T., Mitchell, J. T., & Cutter, S. L. (2010). Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following Hurricane Katrina. *Cartography and Geographic Information Science, 37*, 57–68. https://doi.org/10.1559/152304010790588052.
Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424–440. https://doi.org/10.1109/MSP.2010.938079.
Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceedings of the 17th international conference on world wide web* (pp. 111–120). ACM. https://doi.org/10.1145/1367497.1367513.
U.S. Census Bureau (2018). Building permit survey. https://www:census:gov/construction/bps/.
Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. *Proceedings of the 23rd international conference on machine learning* (pp. 977–984). ACM. https://doi.org/10.1145/1143844.1143967.
Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456). ACM. https://doi.org/10.1145/2020408.2020480.
Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuoustime model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433). ACM. https://doi.org/10.1145/1150402.1150450.
Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of eaq articles from 1965 to 2014. *Educational Administration Quarterly, 53*, 289–323. https://doi.org/10.1177/0013161X16660585.
Riggs, W., Steins, C., & Chavan, A. (2017). City planning department technology benchmarking survey 2017. Technical Report Planetizen https://www:planetizen:com/node/90628/city-planning-department-technology-benchmarking-survey-2017.
Riggs, W., Chavan, A., & Steins, C. (2015). City planning department technology benchmarking survey 2015. Technical Report Planetizen https://www:planetizen:com/node/73480/city-planning-department-technology-benchmarking-survey-2015.
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics, 16*. https://doi.org/10.1186/1471-2105-16-S13-S8 S8.

Data sources

Austin Buildings (2013). *Building footprints.* City of Austin Department of Communications and Technology Management. Retrieved from https://data:austintexas:gov/Locations-and-Maps/Building-Footprints-Year-2013/7bns-7teg.
Austin DOB (2018). *Issued construction permits.* City of Austin Department of Development Services. Retrieved from https://data:austintexas:gov/Building-and-Development/Issued-Construction-Permits/3syk-w9eu/data.
Boston Buildings (2018a). *Boston buildings.* Boston Department of Innovation and Technology. Retrieved from https://data:boston:gov/dataset/boston-buildings1.
Boston DOB (2018). *Approved building permits.* Boston Department of Inspectional Services. Retrieved from https://data:boston:gov/dataset/approved-building-permits.
Chicago Buildings (2018b). *Building footprints.* City of Chicago Department of Buildings. Retrieved from https://data:cityofchicago:org/Buildings/Building-Footprints-current-/hz9b-7nh8.
Chicago DOB (2018). In City of Chicago Department of Buildings (Ed.). *Building permits*. Retrieved from https://data:cityofchicago:org/widgets/ydr8-5enu.
LA Buildings (2018). Countywide building outlines. Retrieved from https://egis3:lacounty:gov/dataportal/2016/11/03/countywide-building-outlines-2014-update-public-domain-release/.
LA DOB (2018). *Building permits.* Los Angeles Department of Building and Safety. Retrieved from https://data:lacity:org/A-Prosperous-City/Building-Permits/nbyu-2ha9.
NYC DOB (2018). *Building job application filings.* New York City Department of Buildings. Retrieved from https://data:cityofnewyork:us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2.
PLUTO (2018). *Primary land use tax lot output (PLUTO)*. New York City Department of City Planning. Retrieved from https://www1:nyc:gov/site/planning/data-maps/open-data/dwn-pluto-mappluto:page.
San Francisco DOB (2018). *Building permits.* San Francisco Department of Building Inspection. Retrieved from https://data:sfgov:org/Housing-and-Buildings/Building-Permits/i98e-djp9.
Seattle SDCI (2018). *Building permits.* Seattle Department of Construction and Inspections. Retrieved from https://data:seattle:gov/Permitting/Building-Permits/76t5-zqzr/data.