Integrating spatial statistics and machine learning to identify relationships between e-commerce and distribution facilities in Texas, US

Kailai Wang[a], Zhenhua Chen[b], Long Cheng[c,d*], Pengyu Zhu[e], Jian Shi[f]

[a] *Department of Construction Management, University of Houston, United States*
[b] *City and Regional Planning, Knowlton School of Architecture, The Ohio State University, United States*
[c] *Jiangsu Key Laboratory of Urban ITS, Southeast University, China*
[d] *Department of Geography, Ghent University, Belgium*
[e] *Division of Public Policy, Hong Kong University of Science and Technology, Hong Kong*
[f] *Department of Engineering Technology, University of Houston, United States*

* *Corresponding author*
E-mail addresses: kwang43@central.uh.edu (K. Wang), chen.7172@osu.edu (Z. Chen), long.cheng@ugent.be (L. Cheng), pengyuzhu@ust.hk (P. Zhu), and jshi23@central.uh.edu (J. Shi)

Abstract

This paper proposes a novel analytical framework that integrates spatial statistics and machine learning to identify relationships between e-commerce and distribution facilities. The framework incorporates centrographic analysis, global and local spatial association measurements, and a recently popularized interpretable machine learning approach – gradient boosting decision trees (GBDT) into warehousing location choice analysis. This framework is applied to the 2003-2016 ZIP Codes Business Patterns data in three large metropolitan areas in Texas, US (i.e., Dallas–Fort Worth, Austin, and Houston). Thematic maps reveal the spatial clustering of areas with more e-commerce activities but less served by logistics facilities. This study does not observe the phenomenon of logistics sprawl occurs in the study region. The GBDT results show the key factors that explain warehousing location choice are industrial activities and transportation network accessibility. The results also suggest, as compared to Dallas-Fort Worth and Austin, the relationship between warehouses and e-commerce establishments is much weaker in Houston – a maritime gateway for goods entering and leaving. Implications for local freight transportation planners and decision-makers are discussed.

Keywords: E-Commerce, Warehouses and Distribution Centers (W&DCs), Spatial Analysis, Logistics Sprawl, Gradient Boosting Decision Trees (GBDT)

1. Introduction

The total US retail e-commerce sales volume has changed dramatically over the last ten years (US Department of Commerce, 2021a). The e-commerce share of total retail sales in US retail increased from 4.2% to 11.3% during 2010-2019. Due to the COVID-19 pandemic, more consumers and businesses have taken to e-commerce almost overnight. In the second quarter of 2020, the percentage reached 15.7% (US Department of Commerce, 2021b). Globally, warehouses are increasing along with e-commerce and supply chain changes. Warehouse location selection has received increasing attention for its potential impacts on local economic activity and the natural and built environment. In the urban setting, logistics and online shopping firms gained customer base and loyalty through fast delivery times and promised delivery windows, such as UPS and FedEx next-day delivery programs and Amazon Prime's promise of two-day shopping. The externalities of freight and logistics activities include greenhouse gas (GHG) emissions, air pollutants, and road traffic congestion and safety concerns (Demir et al., 2015). The purposed study will develop a novel analytical framework and apply it to identify relationships between e-commerce and distribution facilities and present how other factors are associated with warehousing location choice.

Extensive studies have used econometric models to identify, describe, interpret, and predict how key variables of interest are associated with location choices of warehousing or logistics facilities (e.g., Jaller et al., 2017; Yuan, 2019; Kang, 2020a; Sakai et al., 2020; Guerrero et al., 2022). These variables could be categorized as socioeconomic characteristics of local communities, transportation network accessibility, transportation and industrial activities, regional agglomeration effects, and local policy supports. In recent years, social science fields, such as natural language processing, image processing, healthcare information management, and travel demand and traffic accident analysis, have widely adopted advanced data analysis methods to examine statistical relationships among the variables of interest (James et al., 2013; Murdoch et al., 2019). In this line of thought, this study adopts a recently popularized interpretable machine learning (ML) approach – gradient boosting decision trees (GBDT) to seek a better understanding of warehousing location choice in the e-commerce era. This approach has several advantages compared to traditional multiple regression and discrete choice models. For example, it relaxes predefined relationships (e.g., linear, quadratic, and exponential functions). It also captures high-dimensional interactions among explanatory variables (i.e., when two or more processes work together, they produce a synergy effect that is greater than their cumulative effects when used individually), providing more accurate predictions (Friedman, 2001; Ding et al., 2018). Knowing nonlinear and threshold effects of influential variables (e.g., e-commerce facilities, access to transportation infrastructure, and industrial activities) on warehousing location choice is crucial for effective planning implementations, which reveals the costs and benefits across the intervals. Besides, the application of GBDT ranks the relative importance of our variables of interest. Uncovering how substantial e-commerce growth is as compared to other influential factors can provide useful information to local planners and policymakers.

Considering all aspects mentioned above, this study incorporates centrographic analysis, global and local spatial association measurements, and ML approaches into warehousing location choice analysis. The analytical process can be helpful for extracting bivariate associations between warehousing and e-commerce activities from a multi-dimensional perspective. For instance, the centrographic analysis measures and visualizes the spatial movements of facilities' weighted centroids. Thematic maps of local spatial association measurements provide evidence of those dedicated areas for policy provisions. This study conducts empirical analysis using data from 2003 to 2016 for three large metropolitan areas (MSA) in Texas – Dallas–Fort Worth, Austin, and Houston. Research datasets are mainly developed based on ZIP Code Business Patterns (ZBP), American Community Survey (ACS), and TxDOT Open Data Portal.

The key contribution of this study is twofold: First, we adopt a novel analytical framework for analyzing warehousing location choice. Second, to the authors' knowledge, this is one of the first studies using longitudinal data of e-commerce establishments, distribution facilities, and transportation and industrial activities collected in Texas, to reveal relationships between e-commerce and distribution facilities. Logistics sectors contribute substantially to Texas's economy. Most prosperous logistics hubs have a competitive business environment and a sufficient number of warehouses and distribution centers to process, store, and distribute their goods (Finch et al., 2017). However, existing literature has focused on studying warehousing location choice in the cites and metropolitan areas of the Pacific Coast, such as Southern California counties and Seattle, Washington (e.g., Dablanc et al., 2014). Logistics services and freight transportation are more vital to the economy of Texas than that of other inland states. Texas's border with Mexico runs for over 1,200 miles, one of the top US trading partners. There is a business-friendly climate in Texas, as well as urban agglomeration effects that allow freight and logistics industry to thrive there (Beyer, 2021). A longitudinal study of determinants of warehouse location choice for local Texas is not only meaningful to local planners and policy makers on infrastructure investments and provisions, but also verifying whether the current knowledge can be generalized into a different geographical context.

This paper begins with a literature review on the relationship between e-commerce and distribution facilities, factors associated with warehousing location choice, and recent research progress on interpretable machine learning, followed by an explanation of research design and methods. Data is then described. Results are organized into three sections: (1) movement patterns of warehousing and e-commerce activities; (2) bivariate spatial relationship between warehousing and e-commerce activities; and (3) multivariate analysis results. The final section concludes remarks, policy implications, and future research.

2. Literature Review

2.1 E-commerce and warehousing activities

In any supply chain, warehousing serves as an intermediate storage location between two successive stages, which includes receiving, storing, order picking, and shipping (Bartholdi & Hackman, 2014; Gu et al., 2007). Supply chain management considers not only how goods distribution systems should be designed but also how systematic decisions affect the quality of service and logistics costs (Onstein et al., 2019). The evolution of e-commerce has influenced the latter aspect largely and transformed the supply chain ecosystem, bringing manufacturers and consumers together at a deeper level. Every home can become a delivery point in an e-commerce era. To a certain extent, today's warehouses and distribution centers (W&DCs) are designed specifically to meet the needs of online retailers – serving customers directly in the business-to-consumer (B2C) process (Boysen et al., 2019; Onstein et al., 2019; Xiao et al., 2021). Boysen et al. (2019) summarized the new challenges in warehousing management as small orders, wide selections for a broader public, speedy deliveries, and deliveries within the time frame promised. Fulfilling these requirements relies on a higher level of warehousing automation (e.g., automated guided vehicles (AGV)-Assisted Order Picking Systems and Autonomous Robots Moving

Shelves) (Boysen et al., 2019), but also depends on the spatial allocation of warehouses and other logistics facilities (Xiao et al., 2021).

Online shopping and omnichannel distribution are eliminating the space-time constraints of traditional shopping activities. Merchandise, commodities, and services can be delivered anywhere and can be purchased at any time. Under certain circumstances, e-commerce is eco-friendly compared to traditional retail (e.g., Jaller & Pahwa, 2020). Depending on where logistics facilities are located along the supply chain, e-commerce services have different environmental impacts. During 2012-2016, downtown Los Angeles has seen a number of smaller W&DCs operated by retailer giants (e.g., Amazon, Walmart, and Target) (Jaller et al., 2017). This can be explained by the possible benefits from proximity to customers as well as the improved quality of service (Woudsma et al., 2016; Onstein et al., 2019). In the following study, Jaller et al. (2020) suggested the evolution of e-commerce has shifted W&DCs closer to urban centers. Consequently, this trend will result in an increase in the number of trips made by trucks and other vehicles, resulting in increased traffic, more emissions, and greater safety concerns for urban centers. More recently, Rai et al. (2022) introduced the term *"proximity logistics"* to describe the development of logistics facilities in high-demand areas, which are essentially urban, dense and mixed-use. This phenomenon is opposed to the well documented logistics sprawl, namely, *'the spatial concentration of logistics facilities in peripheral regions of metropolitan areas (i.e., moving away from inner urban areas toward more suburban and exurban areas) in a given time period'* (Dablanc & Rakotonarivo, 2010).

2.2 Factors that affect the location choice of warehouses

Existing studies have shown great interest in knowing what factors related to warehousing location choice. These factors can be summarized as socioeconomic characteristics of local communities, transportation network accessibility, transportation and industrial activities, regional agglomeration effects, and local policy supports (e.g., Giuliano et al., 2016; Giuliano & Kang, 2017; Guerrero et al., 2022; Jaller et al., 2017; Jaller et al., 2020; Kang, 2020a; Sakai et al., 2020; Yuan, 2019). There are several research reports published that discuss spatial dynamics of W&DCs in California. Giuliano and her colleagues (2016, 2017) analyzed spatial trends of logistics industry and examined possible explanatory factors. At the descriptive level, W&DCs activity is distributed approximately with population and employment centers. For example, the four largest metropolitan areas in California account for nearly 90% of all W&DCs jobs.

The estimated statistical models in two reports (Giuliano et al., 2016; Giuliano & Kang, 2017) suggest that there was a significant decrease in the correlation between employment density and W&DCs activity. The signs and significances of explanatory variables show that access to transportation infrastructure, such as distances to intermodal terminals, highways, and seaport, and the share of linked industry sectors in the region are important factors to W&DCs activity. Yuan (2019) found that W&DCs are more likely to be located in neighborhoods with higher percentage of minorities in the Los Angeles metropolitan area. However, warehouse developers are not often attracted to low-income neighborhoods due to a lack of convenient amenities, including land availability and transport access.

At a more disaggregated level, Kang (2020a) examined the factors associated with the location choices of more than five thousand warehousing facilities in Los Angeles. This study revealed that facility size and built year influence warehousing location choice significantly. To be more precise, lower land prices and proximity to airports/intermodal terminals are most influential to warehouses built after 2000; while those warehouses built before 1980 are more likely to be influenced by market conditions, labor availability, and proximity to seaports/ intermodal terminals. In another study, using the data from the Paris Region, France, Sakai et al. (2020) analyzed the locational characteristics for multiple types of logistics facilities, including storage facilities operated by logistics service providers, manufacturers, and distributors. The unveiled determinants are consistent with above mentioned studies. This study further indicated the importance of the conducive sociopolitical environments (i.e., zoning policies and land use regulations). Guerrero et al. (2022) incorporated inbound and outbound of freight flows into the analysis of warehousing activities. Yet, little is known about the exact or actual relationship between e-commerce and distribution facilities in different contexts.

The common point of previous studies is that researchers usually rely on econometrics methods for multivariate analysis, such as simultaneous equation model, discrete choice model, censored regression model, and spatial regression. These approaches, however, are restricted to the predefined model structure, and/or allow specific assumptions for data distributions and parameters. Recent developments in the field of interpretable machine learning (ML) offer new opportunities for a fine-grained analysis.

2.3 The application of interpretable machine learning algorithms

ML methods are widely used in transportation and logistics research and exhibit powerful predictive performance. These methods, however, are often criticized for their lack of interpretability, making it difficult to explain the relationships between outcomes and independent variables. Hence, a growing body of research has been conducted to improve the descriptive accuracy and relevancy of machine learning (e.g., Alsaleh & Farooq, 2021; Koushik et al., 2020; Murdoch et al., 2019). A range of model-agnostic methods for interpretation have been proposed, including feature importance, partial dependence, individual conditional expectation (ICE), accumulated local effects (ALE), and localized interpretable model-agnostic explanations (LIME) and Shapley values for local prediction (Molnar, 2022). Cheng et al. (2019) adopted feature importance – represented by the Gini impurity index – to estimate the relative importance of socio-demographics and built environment characteristics on travel outcomes. The average impact of explanatory variables on model predictions can be illustrated by accumulating local effects. Gao et al. (2021) utilized this approach to investigate threshold and interaction effects of different factors on air travel satisfaction of passengers, which yielded easily interpretable analysis results.

Gradient boosting decision trees (GBDT) is one of the most popular interpretable machine learning methods due to its advantages of high prediction accuracy and fast computation (Friedman, 2001). The method has been applied to traffic safety analysis (Tang et al., 2019), travel demand prediction (Ding et al., 2018), and traffic control optimization (Mao et al., 2021). Jin et al. (2022) compared the model performance of GBDT with multiple statistical regressions, such as ordinary least squares, spatial lag model, and spatial error model. They found that GBDT shows highly superior performance – measured by model goodness-of-fit and root mean square error - in estimating how transit accessibility influences housing prices. In the works of Ding et al. (2018) and Yang et al. (2022), GBDT is used to show the existence of nonlinear and threshold relationships between the built environment and travel outcomes, i.e., vehicle miles travelled (VMT) and walking duration. The identified nuanced relationship could offer tailored environmental interventions that benefit sustainable urban mobility visions.

Despite recent research developments discussed above, the bivariate relationship between e-commerce and distribution facilities has not been well explored yet. The present study fills this gap and provides empirical evidence on whether and how e-commerce matters in the occurrence of proximity logistics or logistics sprawl phenomenon. The proposed novel analytical framework in the following section not only measures spatial correlations between e-commerce and distribution facilities, but also reveals nonlinear relationships between warehousing location choice and its well-documented determinants.

3. Research Design and Methods

Figure 1 presents the analytical framework of this study, which consists of three stages – data collection, bivariate statistics with spatial visualization, and multivariate analysis. We create a research dataset based on ZIP Code Business Patterns (ZBP), TxDOT Open Data Portal, and American Community Survey (ACS). The dataset contains information on warehousing and e-commerce activities, transportation activities, transportation network accessibility, and relevant socioeconomic factors at the zip code level. Then we extract bivariate associations between warehousing and e-commerce activities from a multi-dimensional perspective. Later, multiple regression model and gradient boosting decision trees (GBDT) model are built to support fine-grained investigation for the relationship between warehousing and e-commerce activities, and identify how other key variables influence warehousing location choice. The main research findings and policy implications are derived based on three sequential phases of analysis.

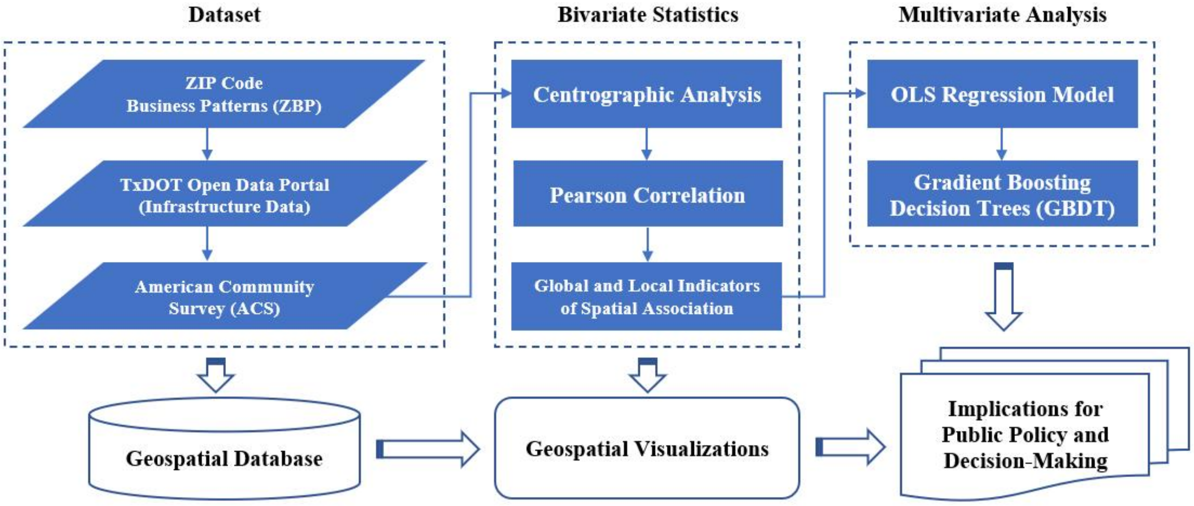

Dataset
Bivariate Statistics
Multivariate Analysis
ZIP Code Business Patterns (ZBP)
TxDOT Open Data Portal (Infrastructure Data)
American Community Survey (ACS)
Centrographic Analysis
Pearson Correlation
Global and Local Indicators of Spatial Association
OLS Regression Model
Gradient Boosting Decision Trees (GBDT)
Geospatial Database
Geospatial Visualizations
Implications for Public Policy and Decision-Making

Figure 1. The workflow of the research protocol

3.1 Centrographic analysis

Centrographic analysis has been commonly used and accepted as an effective tool to investigate the spatial movements of warehouses or logistics facilities (e.g., Dablanc et al., 2014; Jaller et al., 2017; Guerin et al., 2021). In this study, centrographic analysis is conducted for the number of warehousing and distribution centers (W&DCs) and e-commerce establishments. We calculate the weighted centroid of facilities' locations each year, and then measures the distances from these weighted centroids to the zip code population weighted centroids. The weighted centroid of a certain type of facilities (i.e., warehouses or e-commerce establishments) is calculated as follows (Yeates, 1974):

$$\bar{x}w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}, \qquad (1)$$

$$\bar{y}w = \frac{\sum_{i=1}^{n} y_i w_i}{\sum_{i=1}^{n} w_i}, \qquad (2)$$

where $\bar{x}w$ latitude coordinate of the weighted centroid in a given year; $\bar{y}w$ longitude coordinate of the weighted centroid in a given year; x_i latitude coordinate of facility i (zip code); y_i longitude coordinate of facility i (zip code); w_i number of facilities within a particular zip code. To track spatial movements, this study visualizes the spatial locations of weighted centroids of W&DCs and e-commerce establishments from the year 2003 to 2016. Also, the distances between the facilities' weighted centroids and the population weighted centroids are calculated, which can provide helpful information on whether our facilities of interest were moving away from or moving to urban centers during the study period.

3.2 Bivariate local indicator of spatial association (LISA)

As this study cannot access to the detailed spatial coordinates of industrial establishments, we estimate the spatial correlations between the numbers of W&DCs and e-commerce for adjacent spatial units. A bivariate local indicator of spatial association (LISA) statistic is implemented (Anselin et al., 2010). Bivariate LISA is a local Moran's *I* metric that intends to uncover the relationships between two spatial factors. Below are the equations for calculating global and local bivariate Moran*'s I* statistic:

$$I^{e-comm\_whs} = \left(\frac{N}{\sum_i \sum_j w_{ij}}\right)\left(\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(A_j - \bar{A})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_j (A_j - \bar{A})^2}}\right), \qquad (3)$$

$$I_i^{e-comm\_whs'} = \frac{N}{\sum_i (x_i - \bar{x})^2}\left(x_i - \bar{x}\right)\sum_j w_{ij}\left(A_j - \bar{A}\right), \qquad (4)$$

where I^{e-comm_whs} and $I_i^{e-comm\_whs'}$ are the global and local bivariate Moran's *I* for e-commerce and warehousing activities, respectively; *N* is the total number of spatial units; w_{ij} is the queen contiguity spatial weight matrix to explore the spatial relationship between unit *i* and unit *j*; x_i is the number of e-commerce establishments of the unit *i*; $\bar{x}$ is the average value of the number of e-commerce establishments in the study area (i.e., one Metropolitan Statistical Area); A_j is the number of W&DCs of the unit *j*; and $\bar{A}$ is the average value of the number of W&DCs in the study area. The value of I^{e-comm_whs} or $I_i^{e-comm\_whs'}$ is between -1 and 1. If one spatial unit having a large number of e-commerce establishments is surrounded by spatial units with adequate W&DCs, then the local bivariate Moran's *I* will return to a postive value. The opposite spatial pattern is represented by a negative value. The magnitude of spatial dependence can be reflected in the estimated value – the stronger spatial correlation, the larger absolute value. To determine whether Moran's *I* value is statistically significant, a permutation test is applied, and the pseudo-significance level is set at the 5% level with 999 permutations (Anselin et al., 2010). This study visualizes local spatial correlations between the number of e-commerce establishments at a specific unit and the mean value of the number of W&DCs at all neighboring units by using cluster maps. There are five categories of local spatial correlations: High-High cluster ($H_{e-comm}H_{whs}$), High-Low cluster ($H_{e-comm}L_{whs}$), Low-High cluster ($L_{e-comm}H_{whs}$), Low-Low cluster ($L_{e-comm}L_{whs}$), and not statistically significant.

3.3 Gradient Boosting Decision Trees (GBDT)

A multiple linear regression model is first estimated for multivariate analysis. Based on the results, we can gain a preliminary understanding of the factors affecting warehousing location choice. However, the linearity assumption has been criticized for resulting in severely biased estimates. A certain explanatory variable's relationship with the number of W&DCs will change over the full range of values. In some areas, businesses and customers are less willing to shift massively to digital marketing services. Inventory management and timely delivery of goods may be overlooked by communities with a lower level of digital maturity. Therefore, the connection between e-commerce and distribution facilities in these areas is weaker than others having more e-commerce fulfillment services. This study uses a recently popular machine learning method – gradient boosting decision trees (GBDT).

Friedman (2001) and many other studies adopted this approach have documented the details of the GBDT algorithm in several intuitive ways (Ding et al., 2018; Dong et al., 2019; Jin et al.,2022; Mao et al., 2021; Tang et al., 2019; Tao et al., 2020; Wang & Ozbilen, 2020; Yang et al., 2022; Zhang & Haghani, 2015). As suggested by its name, there are several single decision trees that are merged together to reach the results. A feature of this algorithm is that it automatically captures the interactions between predictors. Each successive model in gradient boosting attempts to predict the error left over by the previous model based on the error left over by the previous model. By summing up all the decision tree results, we can predict the outcome.

Relative importance and partial dependence of explanatory variables are often used to interpret the “black-box” model structure. The relative importance of explanatory variable x_i is calculated as follows:

$$I_{x_i}^2 = \frac{1}{M}\sum_{m=1}^{M} I_{x_i}^2(T_m), \qquad (5)$$

$$I_{x_i}^2(T_m) = \sum_{j=1}^{J-1} d_j \, \{split\ at\ node\ j\ is\ on\ variable\ x_i\}, \qquad (6)$$

where *J* is the number of leaves on each tree; T_m is the *m^th^* tree function;d_j represents the improvement in the squared error by making the *j^th^* split using based on predictor x_i. The relative importance of all explanatory variables adds up to 100%.

Partial dependence plots show marginal effects of our variables of interest on the predicted response variable. Mathematically, the partial dependence of $F(x)$ on x_s can be formulated as follows:

$$F_{x_s}(x_s) = \ E_{x_c}[F(x_s, x_c)] = \int F(x_s, x_c) dP(x_c)\ , \qquad (7)$$

$$\overline{F}_{x_s}(x_s) = \ \frac{1}{n}\sum_{i=1}^{n} F(x_s, x_{ic}), \qquad (8)$$

where x_s are the features whose specific effects on the predicter response variable are to be estimated; x_c are other explanatory variables; $P(x_c)$ is the probability density function of x_c; *n* represents the sample of model estimation. In partial dependence, the model output is marginalized over the distribution of these other predictors. Partial dependence plots also include interactions between predictors due to the model's ability to handle interaction effects among predictors.

4. Data and Study Area

This study performs spatial analyses for e-commerce activity and locations of warehousing and distribution centers (W&DCs) in Texas. the ZIP Code Business Pattern (ZBP) database contains the measurements of two types of activities. This database provides information about the number of establishments of each industrial classification (NAICS) at the zip code level. In its classification system, the NAICS uses a six-digit coding system to categorize different types of industries. W&DCs are classified under NAICS 493. In this study, e-commerce activity refers to the number of establishments primarily engaged in retailing all types of merchandise using non-store means, in terms of Electronic Shopping and Mail-Order Houses, which are under NAICS 454110. Over the past decade, the evolution of e-commerce has led to a surge in demand for warehousing and logistics facilities in Texas. Amazon, FedEx and Lowe's are three of the companies that have built many new fulfillment centers and delivery stations in response to the accelerating e-commerce growth (Mahoney, 2020; Thomas, 2021). Figure 2 shows the spatial distributions of W&DCs and e-commerce establishments based on the 2016 ZBP database. Clearly, more extensive warehousing and e-commerce activities are located in the Houston-The Woodlands-Sugar Land Metropolitan Statistical Area (Houston), the Dallas-Fort Worth-Arlington Metropolitan Statistical Area (Dallas–Fort Worth), and the Austin–Round Rock-Georgetown Metropolitan Statistical Area (Austin) than the rest of Texas. The analyses below thereby focus on the three selected MSAs.

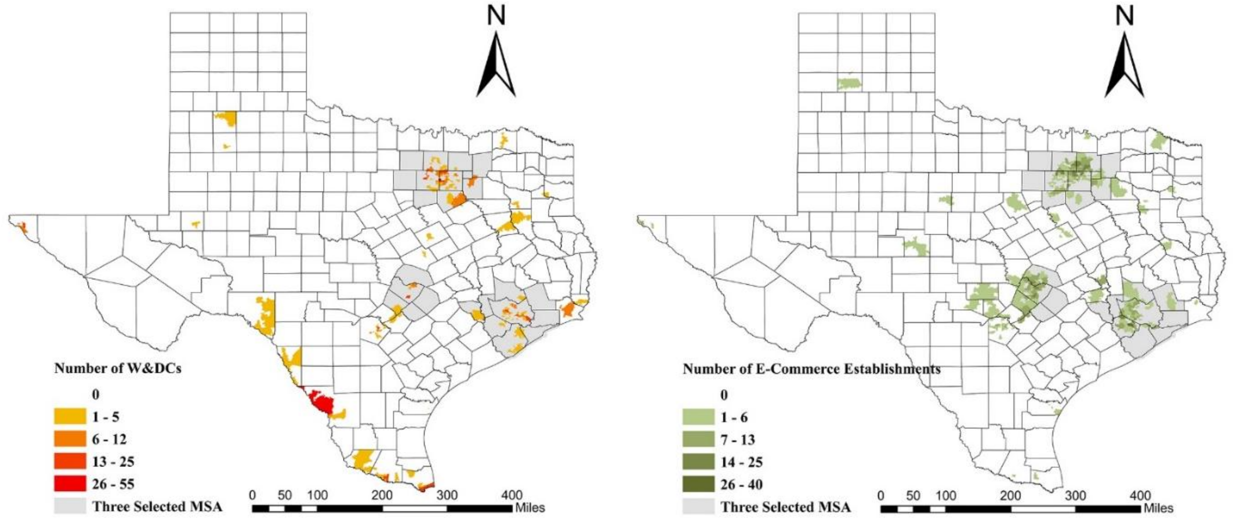


Figure 2. Spatial distribution of W&DCs and e-commerce establishments in 2016 in Texas

We build the research dataset that covering the time period 2003-2016 for two reasons. First, the North American Industry Classification System (NAICS) is updated every five years to keep up with changes in economics. Before 2017, many mini-storage businesses are classified under NAICS 493 Warehousing and Storage. These establishments are NAICS 531130 Lessors of Miniwarehouses and Self-Storage Units in 2017 and afterwards (Woudsma et al., 2016; Woudsma & Jakubicek, 2020). The change will result in data consistency if we combine the pre-

2017 and post-2017 datasets together. Second, the COVID-19 pandemic has caused supply chain disruptions since the beginning of 2020. When the time period before the pandemic is considered, empirical evidence can be obtained under normal circumstances, which removes the COVID-19 impact on the analytical results.

The logistics businesses are expected to have the most immediate connection with warehousing business (Kang, 2020b). This study thus considers the numbers of establishments in air transport, water transport, and truck transport from the ZBP database. These transportation activities are under NAICS 481, 483, and 484. As consumers, shippers, and receivers of freight shipment, manufacturing, wholesale, and retail trade sectors can be involved in goods distribution. The intensities of these sectors can provide evidence for the diversity of industrial activities. Recent studies have revealed a direct link between these sectors and warehouse operations. The proximity to manufacturing and retail facilities means being close to distribution channels, transportation, and relevant infrastructure (Yuan, 2019; Jaller et al., 2017; Kang, 2020a). To quantify the intensity, this study divides the number of establishments of each industrial sector by the land area at the zip code level. The number of establishments engaged in manufacturing, wholesale, and retail industries are under NAICS 31, 42, and 44, respectively. The land area information comes from the 2010 US Census.

Other variables of our interest are also measured at the zip code level. For the proximity to customers and transportation networks, the nearest distances between the centroid of each spatial unit and four types of transportation infrastructure are calculated, in terms of highways, seaports, airports, intermodal facilities. The calculations are based on the TxDOT GIS Open Data Portal[1]. The 2015-2019 American Community Survey (ACS) 5-year Estimates offers the most recent socioeconomic factors. This study adopts median household income, race and ethnicity groups, housing occupancy status, and household internet subscriptions as explanatory variables. The data on population density comes from the 2010 US Census.

There are 1,675 ZIP Code Tabulation Areas (ZCTAs) in Texas that have complete information on W&DCs, e-commerce establishments, transportation activities, intensities of industrial activities, and access to transportation infrastructure at each year during the study period. We have 23,450 observations in total. However, in some ZCTAs, socioeconomic factors provided by the 2015-2019 ACS for certain years are missing. Those observations were excluded from the multivariate analyses. Finally, we have 22,554 valid observations. Table 1 summarizes the descriptive statistics of final sample characteristics for four geographic contexts.

[1] 1) Highway: https://gis-txdot.opendata.arcgis.com/datasets/txdot-texas-highway-freight-network/explore?location=31.139063%2C-100.049294%2C6.69;
2) Seaport: https://gis-txdot.opendata.arcgis.com/datasets/txdot-seaports/explore?location=28.021200%2C-95.665187%2C8.05;
3) Airport: https://gis-txdot.opendata.arcgis.com/datasets/texas-airports/explore?location=31.173825%2C-100.059833%2C6.72; 4) Intermodal Facility: https://gis-txdot.opendata.arcgis.com/datasets/txdot-national-highway-system/explore?location=30.986103%2C-100.084847%2C6.69

Table 1. Data Summary

Descriptive Statistics of Variables	Dallas		Austin		Houston		Others	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ZIP Code Business Patterns								
Warehousing and Distribution Centers (W&DCs)	1.29	2.94	0.56	1.34	1.08	2.24	0.47	1.83
E-Commerce Establishments	1.67	2.68	2.32	3.22	1.24	2.11	0.29	0.77
Industrial activities								
Manufacturing intensity (establishments/mile2)	2.02	4.55	1.17	2.09	2.65	5.02	0.45	2.49
Wholesale trade intensity (establishments/mile2)	4.03	11.20	1.89	3.73	5.55	15.16	0.95	7.48
Retail trade intensity (establishments/mile2)	8.19	19.14	6.12	11.71	8.97	15.40	2.54	12.86
Transportation activities								
Air transport (establishments)	0.42	2.25	0.22	0.88	0.44	1.95	0.15	0.73
Water transport (establishments)	0.04	0.22	0.02	0.15	0.35	0.89	0.03	0.19
Truck transport (establishments)	6.29	6.99	3.75	4.19	5.71	6.33	4.25	9.80
TxDOT GIS Open Data Portal								
Access to transportation infrastructure								
Distance to highway (mile)	1.56	2.22	1.52	1.90	1.57	1.98	3.88	5.70
Distance to seaport (mile)	230.83	20.37	120.46	17.92	29.55	18.30	214.49	166.13
Distance to airport (mile)	6.54	3.73	8.65	4.14	7.59	3.91	10.59	6.99
Distance to intermodal facility (mile)	12.55	13.13	76.64	15.70	8.90	10.77	52.23	35.81
2015-2019 American Community Survey 5-year Estimates								
Median household income (in $1000)	71.48	28.46	79.70	29.45	69.20	31.67	52.84	17.77
Race								
% of population is white	73.5%		79.0%		69.8%		84.9%	
% of population is black	13.4%		6.0%		16.6%		6.8%	
% of population is asian	4.8%		4.1%		4.9%		0.9%	
% of population is other races (including American Indian and Alaska Native alone, Native Hawaiian and Other Pacific Islander alone, and others)	8.3%		10.9%		8.7%		7.4%	
Ethnicity								
% of population is non-Hispanic	74.4%		69.3%		66.3%		64.8%	
% of population is Hispanic	25.6%		30.7%		33.7%		35.2%	
Housing occupancy status								
% of occupied housing units	90.8%		89.0%		87.3%		77.8%	
% of vacant housing units	9.2%		11.0%		12.7%		22.2%	
Internet subscriptions in household								
% of households have internet subscription	83.5%		84.8%		81.7%		73.0%	
% of households have internet access without a subscription	2.9%		2.7%		2.5%		3.7%	
% of households do not have internet access	13.6%		12.5%		15.8%		23.4%	
2010 US Census								
Population density (1000*persons/mile2)	2.10	2.45	1.45	2.14	2.36	2.44	0.51	1.20
Number of observations before combing ACS data	4046		1316		3360		14728	
Number of observations without missing	4018		1316		3234		13986	

5. Results

5.1 Spatial movement patterns of warehousing and e-commerce activities

Table 2 Changes in the number of warehouses

Year	Dallas	Austin	Houston	Others
2003	316	39	212	386
2004	342	48	233	407
2005	345	47	226	427
2006	362	54	237	440
2007	375	60	248	467
2008	387	55	247	502
2009	375	52	248	518
2010	364	51	258	510
2011	345	51	250	535
2012	367	47	277	510
2013	387	55	287	528
2014	405	57	290	519
2015	421	63	301	543
2016	437	63	315	562

Table 2 shows the changes in the number of warehouses in three MSAs and the rest of Texas during 2003-2016. Roughly speaking, we observe a relatively even rise before 2008 and after 2012. The fluctuations occurred during 2008-2012 can be because of the disruptions caused by the economic crisis. The number of W&DCs and the number of e-commerce establishments in 2003, 2010, and 2016 for three MSAs are visualized in Figure A1 (in Appendix). For example, the number of W&DCs experienced a more substantial increase in the central part of the Dallas–Fort Worth than other areas in this MSA. As to the number of e-commerce establishments, a more noticeable growth occurred in the northeast part of Dallas–Fort Worth. In Austin, it seems that the spatial patterns of W&DCs and e-commerce establishments were changing synchronously. More increases took place in the west part, including Travis, Hays, and Williamson counties. In Houston, the number of W&DCs increased significantly along the highways I-10 and I-69. This increment synchronized with the trajectory of urban growth. The number of e-commerce establishments increased across the Houston region, having a greater concentration in the central part. Further analyses are needed to uncover the factors that drive the changes in spatial patterns.

Figure 3 displays the yearly weighted geometric centers of W&DCs and e-commerce establishments in three MSAs during 2003-2016. We infer that more e-commerce establishments were built in the northern part of Dallas–Fort Worth than in the southern part. Clearly, the weighted centroids of e-commerce establishments are moving to the western part of Houston. In the Austin region, the weighted centroids were slightly shifting towards the east.

Regarding spatial changes in warehousing activities, as shown in Figure 3, there was no apparent change of the weighted centroids in Dallas–Fort Worth and Houston during the study period. In Dallas–Fort Worth, these weighted centroids were in close proximity to the Texas 183 TEXpress in the City of Irving, which is about 10 miles northwest of downtown Dallas. While in Houston most centroids were within or close to the downtown area. Somewhat interestingly, we spot a clear trace of the weighted geometric centers of W&DCs moving from the north to the south of the Austin region. Austin is a fast-growing metropolitan area whose percentage of urbanized land has increased significantly in the south side since 2006 (Guo & Zhang, 2021). In sum, the trajectories of weighted centroids during 2003-2016 do not support the existence of a significant sprawl in logistics facilities in Texas. The results are in line with the literature. Cidell (2010) reported that during 1986-2005, freight activities are more concentrated in central counties in Texas cities, which is different from other US metropolitan areas. This can be attributed to municipal policies that either ‘actively stimulate’ or discourage logistics activities. The distances between warehousing centroids and central cities of three MSAs (i.e., 2010 US census population-weighted centroids) on a yearly basis are displayed in Figure 4.

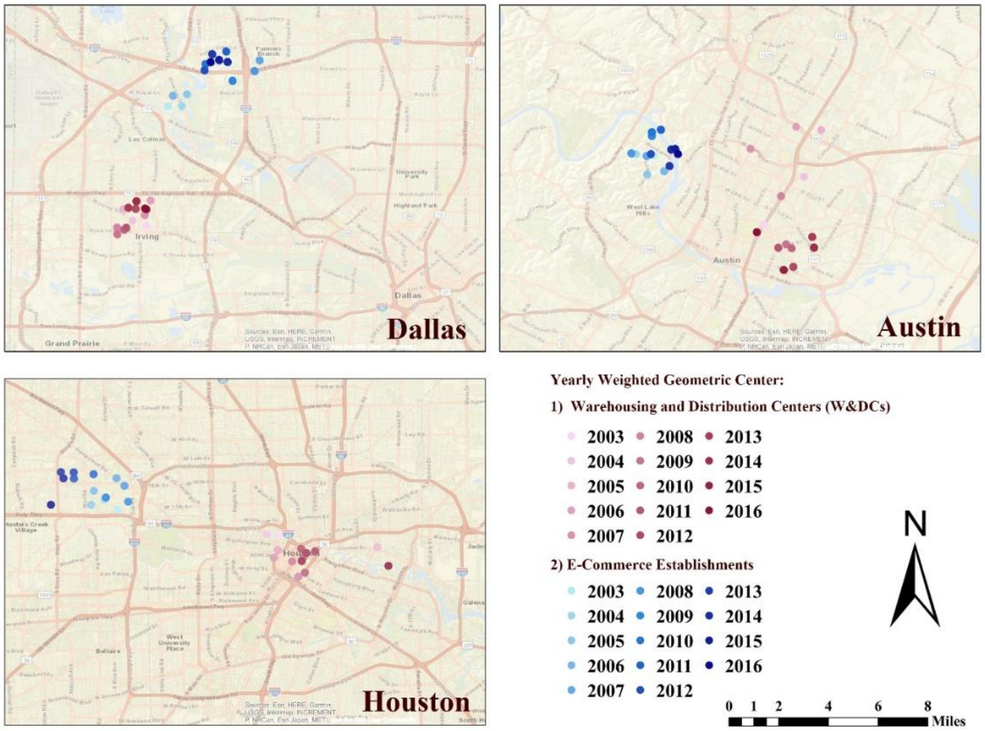


Figure 3. The weighted geometric centers of warehousing and distribution centers (W&DCs) and e-commerce establishments in three MSAs during 2003-2016

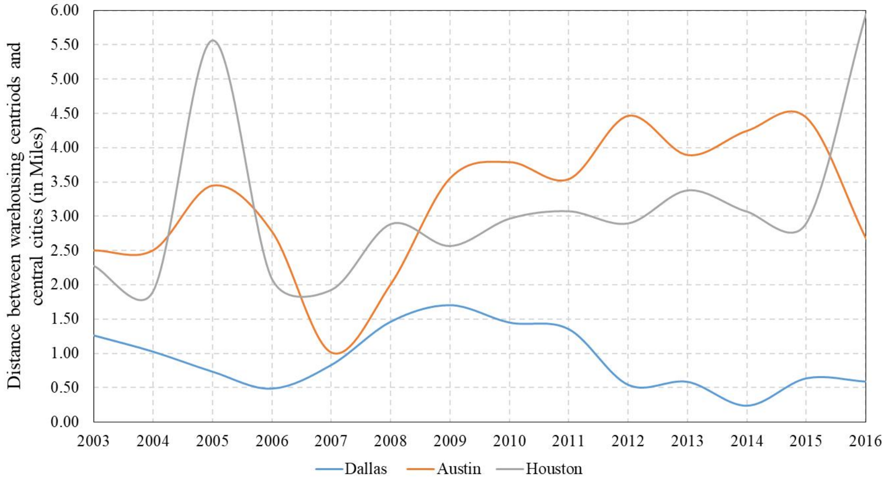


Figure 4. Distances between warehousing centroids and central cities of three MSAs

5.2 Bivariate association between warehousing and e-commerce activities

This study performs the correlation analysis for the key variables of interest in two dimensions: for non-spatial relationships, Pearson correlations are used, while for neighbor dependencies, bivariate spatial correlations are used. Table 3 presents the Pearson correlations between the numbers of W&DCs and e-commerce establishments for all the years of research. The scales of the estimated coefficients suggest that the correlations between W&DCs and e-commerce establishments had varied over the years and across three geographic contexts. There is a much closer relationship between e-commerce and distribution facilities in Dallas–Fort Worth and Austin than in the Houston region.

Table 3. Pearson correlations between number of W&DCs and e-commerce establishments

	Dallas		Austin		Houston	
Year	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.
2003	0.3940	**	0.2473	**	0.1144	*
2004	0.4044	**	0.3111	**	0.1200	*
2005	0.4306	**	0.3676	**	0.0904	
2006	0.4559	**	0.4355	**	0.0595	
2007	0.3211	**	0.4407	**	0.0635	
2008	0.3145	**	0.4695	**	0.0682	
2009	0.3094	**	0.4907	**	0.0562	
2010	0.3283	**	0.4484	**	0.0346	
2011	0.3111	**	0.4477	**	0.1532	**
2012	0.3694	**	0.2800	**	0.1164	*
2013	0.3488	**	0.2586	**	0.1382	**
2014	0.3439	**	0.2024	*	0.1507	**
2015	0.3386	**	0.2089	**	0.1249	*
2016	0.3318	**	0.2122	**	0.1392	**

Notes: ∗∗ Significant at the 95% level; ∗ Significant at the 90% level.

The global bivariate Moran's I reported in Figure 5 suggest that geographical correlations between W&DCs and e-commerce establishments are positive in three MSAs at the 2003, 2010, and 2016. Somewhat surprisingly, we find that in Austin and Houston, the correlation was weaker in 2016 than in 2003. The bivariate local indicator of spatial association (LISA) map illustrates four kinds of spatial autocorrelations between warehouse and e-commerce activities. Notably, the Low-High cluster refers to the areas with low intensity of e-commerce activities but high intensity of warehousing activities, while the High-Low cluster refers to the inverse ones. Most positive spatially matched areas (High-High clusters) are in urban central areas or dense urban environments in three MSAs, suggesting these communities with high e-tailing delivery demand are well served by logistics facilities (i.e., the number of W&DCs). In Houston and Dallas–Fort Worth, the negative counterparts (Low-Low clusters) are mainly located in urban peripheries. Particularly, we observe a substantial increase in the number of Low-Low clusters located in urban peripheries of Dallas–Fort Worth during 2003-2016. Perhaps, these areas have a low demand of e-tailing delivery services and do not have too many logistics facilities. There also exists an obvious spatial mismatch. The Low-High clusters representing less e-commerce activities, but more warehouse supplies, are more likely to locate in central urban areas and close to High-High clusters. The number of High-Low clusters is decreasing during 2003-2016 in Houston and Dallas–Fort Worth. These areas have more e-commerce activities but are less served by logistics facilities (i.e., the number of W&DCs). As the market demand of rapiddelivery programs is continuously increasing, these areas are more likely to be the underserved areas that need particular policy interventions and provisions.

5.3 Factors associated with warehousing location choice

The purpose of estimating regression models is to identify the factors that are significantly associated with warehousing location choice and provide initial assessments of their influential directions. The details are available in Table A1 (in Appendix). In the presence of multicollinearity, the results of relative importance ranking, and partial dependence plots may be biased. To address this concern, we report VIF values for all explanatory variables in all statistical models, and find no multicollinearity exists. As expected, the number of e-commerce establishments is positively associated with the number of W&DCs. The estimated margins suggest that, after controlling for confounding factors, the correlation between e-commerce and warehousing activities in Dallas–Fort Worth is stronger than that in the rest of the state. Consistent with previous studies, the regression results show that transportation activities, transport network accessibility, second-order industrial activities, and other socioeconomic factors significantly affect warehousing activities. For example, warehouses are more likely to be in the neighborhoods with better transportation infrastructure. As it has been reemphasized by many new economic geography models, transportation cost is a key factor that affects the location choice of individuals and firms from the economic perspective.

Table 4 presents the relative importance of independent variables that influence warehousing location choice in three MSAs. The higher the relative importance, the more contribution a variable makes to the prediction. E-commerce facilities contributes to 5.40% (ranking 4th) and 6.23% (ranking 5th) in the models for Dallas–Fort Worth and Austin, respectively. However, this variable seems much less pronounced in the model for Houston (1.46% and ranking 18th). The results do not support spatial coincidence between warehousing and e-commerce activities occurred in Houston during the study period. Container logistics and maritime transport may largely influence warehousing location choice in Houston. Figure 6(a) illustrates how the number of e-commerce establishment influences warehousing activities in three MSAs differently. Nonlinearities and threshold responses do exist. For example, in Dallas–Fort Worth, the number W&DCs increases sharply in a nearly-linear pattern within the range of 5-9 e-commerce establishments. The influences of e-commerce on warehousing activities become trivial outside the range abovementioned. In Austin and Houston, the influence is not as visible as that in Dallas–Fort Worth. In order to better understand nonlinear relationships between e-commerce and distribution facilities, future studies should make use of data on online purchases that retailers can gather and process more easily, whereas researchers have limited access.

Collectively, access to transportation infrastructure, industrial activities, and race and ethnicity groups are the most influential components determining the number of W&DCs in all three models. The sum of three components is 62.44%, 73.82%, and 72.24% in the models for Dallas–Fort Worth, Austin, and Houston, respectively. Considering the importance of two variables representing industrial activities in all three models, we further provide partial dependence plots. The relative contribution of manufacturing intensity to the variation in predicting the number of W&DCs in models for Dallas–Fort Worth, Austin, and Houston is 18.53% (ranking 1st), 15.05% (ranking 2nd), and 21.55% (ranking 1st), respectively. Figure 6(b) shows the margins of manufacturing intensity in three MSAs. An increment in manufacturing intensity has a stronger positive impact on the number of W&DCs in Dallas–Fort Worth than that in Austin and Houston. This is quite similar to what we found for e-commerce establishments. While the number of W&DCs in Austin is more sensitive to the changes in manufacturing intensity than that in Houston.

Based on the reported relative contributions in Table 4, we infer that retails trade intensity also influence warehousing activities largely. The margins plotted in Figure 6(c) show that the influences of this factor fluctuate drastically across the whole observed range for all three regions. In other words, the relationships between retail trade intensity and the number of W&DCs vary greatly at different intervals. Further research is needed to show the mechanism of the fluctuation.

Among access to transportation infrastructure variables, we found that the distance to seaport has a relative larger contribution of 10.54% (ranking 3rd) in the Houston model than in models for Dallas–Fort Worth and Austin. This can be because many warehousing and logistics activities in the Houston region are related to maritime transportation. Figure 2 illustrates a cluster of W&DCs around the Port of the Houston to southeast. This could also be the reason access to intermodal facility is a more substantial factor (7.01% and ranking 5th) in Houston. The intermodal facility gathers different modes of transportation and is strategically located to improve regional mobility.

Race and ethnicity groups are much more influential predictors of the number of W&DCs in Austin as compared to those in Dallas–Fort Worth and Houston. Particularly, the share of other races population, in terms of American Indian and Alaska Native alone, Native Hawaiian and Other Pacific Islander alone, and others, contributes the greatest (15.32% and ranking 1st) in the Austin model among all single variables evaluated. The percentages of Black population and Hispanic reaches 7.08% (ranking 4th) and 6.11% (ranking 6th) in the Austin model, respectively. While workers of color make up 37% of the US labor force, they comprise more than two thirds of workers in the warehousing industry and more than half in e-commerce[2]. The research results reveal that workers of color are likely to play more important roles in logistics industry in Austin than the other two studied MSAs.

The year-specific variable explains 2.54% (ranking 16th), 5.35% (ranking 8th), and 3.63% (ranking 9th) of the variations in models for Dallas–Fort Worth, Austin, and Houston, respectively. This trend variable can be seen as a proxy for inventory management practices, warehouse automation solutions, and others that influence warehousing activities and are highly correlated with time. However, they are not directly observable for this study. Figure 6(d) visualizes the margins of the year-specific variable for three regions. Overall, there exists a gentle and nearly-linear increase before 2007 or 2008, and then decreases slightly during 2008-2012. After 2012, the number of W&DCs increases in all three regions as time goes by.

[2] Link: https://laborcenter.berkeley.edu/pdf/2019/Future-of-Warehouse-Work.pdf

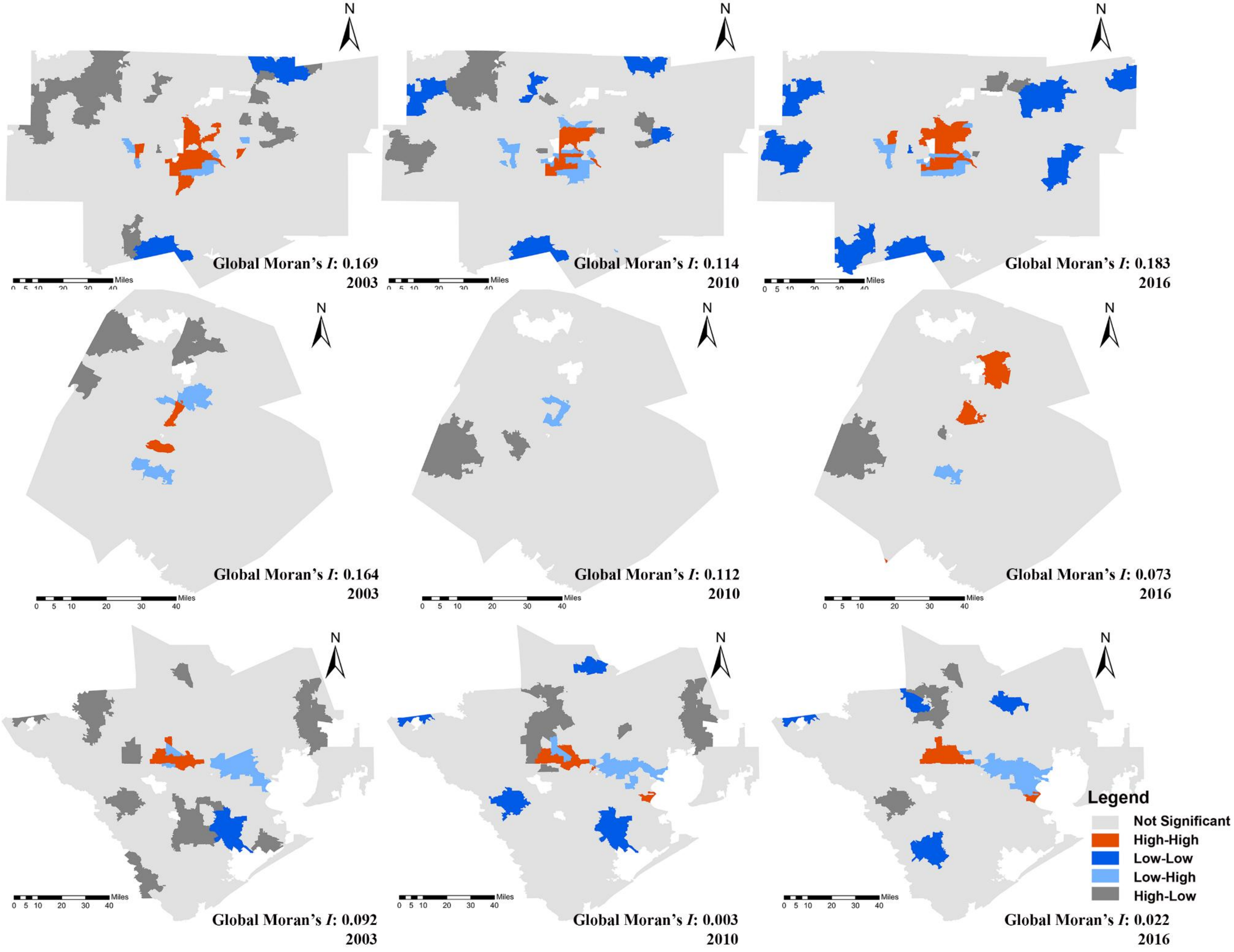
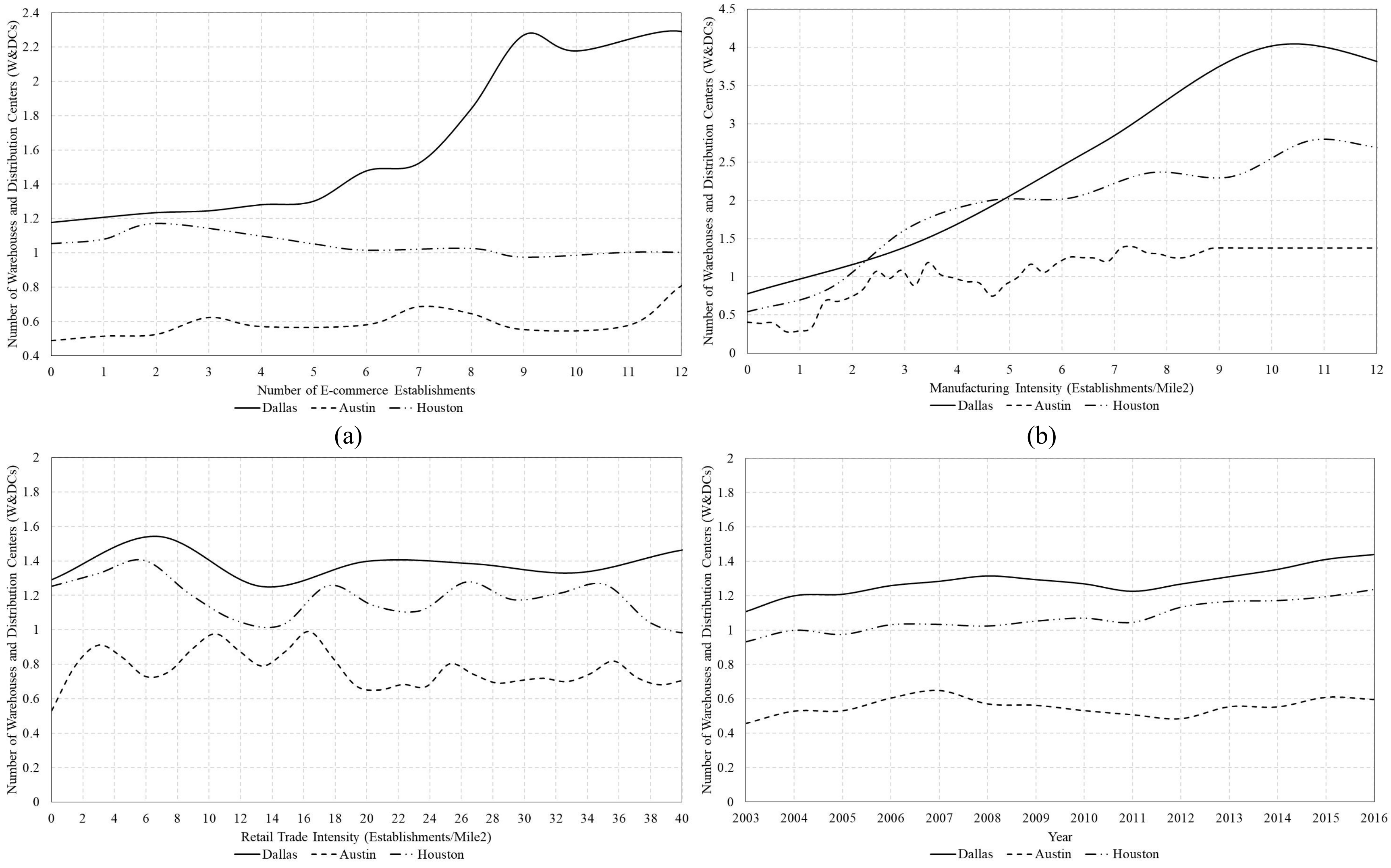

N
Global Moran's I: 0.169
2003
Global Moran's I: 0.114
2010
Global Moran's I: 0.183
2016
Global Moran's I: 0.164
2003
Global Moran's I: 0.112
2010
Global Moran's I: 0.073
2016
Global Moran's I: 0.092
2003
Global Moran's I: 0.003
2010
Global Moran's I: 0.022
2016
0 5 10 20 30 40 Miles
Legend
Not Significant
High-High
Low-Low
Low-High
High-Low

Figure 5. The spatial cluster maps between warehouse and e-commerce activities in 2003, 2010, and 2016

Table 4. Relative Contributions of Independent Variables on Warehousing Activities

	Dallas		Austin		Houston	
	Rel. Imp. (%)	Rank	Rel. Imp. (%)	Rank	Rel. Imp. (%)	Rank
E-Commerce Facilities						
Number of e-commerce establishments	5.40	6	6.23	5	1.46	18
Transportation activities	5.46		1.07		2.19	
Number of air transport establishments	3.79	11	1.07	18	0.34	19
Number of water transport establishments	1.67	19	0.00	19	1.85	15
Access to transportation infrastructure						
Distance to highway	6.34	4	5.85	7	2.57	12
Distance to seaport	2.34	17	3.10	12	10.54	3
Distance to airport	4.72	8	3.07	13	4.94	7
Distance to intermodal facility	3.80	10	2.71	15	7.01	5
Industrial activities						
Manufacturing intensity	18.53	1	15.05	2	21.55	1
Retail trade intensity	5.89	5	12.19	3	10.49	4
Population density	9.40	3	2.74	14	11.74	2
Median household income (in $1000)	5.10	7	4.42	9	3.55	10
Race (base case: % of population is white)						
% of population is Black	3.55	13	7.08	4	2.28	13
% of population is Asian	3.59	12	3.34	11	3.18	11
% of population is other races	2.27	18	15.32	1	4.34	8
Ethnicity						
% of population is Hispanic	11.41	2	6.11	6	5.33	6
Housing occupancy status						
% of vacant housing units	4.38	9	3.37	10	1.59	17
Internet subscriptions in household						
% of households have internet subscription	2.69	14	1.60	16	1.69	16
% of households have internet access without a subscription	2.58	15	1.38	17	1.92	14
Year	2.54	16	5.35	8	3.63	9
Tuning Parameters and Model Fitness						
Number of trees	24334		8244		13334	
Number of leaves (tree complexity)	20		20		20	
Shrinkage (learning rate)	0.01		0.01		0.01	
Minimum number of samples in terminal nodes	15		15		15	
Subsampling fraction	0.5		0.5		0.5	
RMSE	0.672		0.440		0.577	
Cross-Validated R-Squared	0.948		0.893		0.934	



(a)

(b)

(c)

(d)

Figure 6. The effects of key variables on the number of warehousing and distribution centers (W&DCs)

6. Conclusion and Discussion

The objective of the study is to propose a novel analytical framework to examine the relationship e-commerce and distribution facilities. Many studies have discussed the phenomenon of logistics sprawl and factors associated with warehousing location choice; however, the number of empirical studies with specific focus on the aforementioned relationship is still relatively small. To demonstrate the use of the framework, we construct the research dataset mainly based on 2003-2016 ZIP Code Business Patterns in Texas, US. This study first carries out a centrographic analysis to reveal spatial movements of warehousing and distribution centers (W&DCs) and e-commerce establishments during the study period. Then the global and local bivariate Moran's I for e-commerce and warehousing activities are calculated and spatially visualized. From the perspective of policy and practices, the revealed results are informative for identifying where e-commerce activity was underserved or oversupplied by warehouses. Finally, this study applies machine learning approach to explore the relative importance and threshold effects of e-commerce establishments and other key factors on the number of W&DCs.

In general, the research results showed that the use of the proposed framework can effectively provide useful insights to local policy makers, logistics service providers, and other practitioners. The visualization of descriptive and explanatory spatial statistics makes the interpretation of how e-commerce and warehousing activities occur more straightforward and comprehensive. The case study also shows the ability of the framework to be conducted at any scale and location to identify relationships between e-commerce and distribution facilities. Below we summarize other important findings and their relevant implications, as well as the possible directions for future research.

The impacts of logistics sprawl have been studied scientifically over the past decade. This study does not find solid evidence that warehouses are sprawling significantly in major metropolitan areas in Texas. Large metropolitan areas suffer from logistics sprawl as they usually act as hubs for export and import operations as well as massive consumer markets. In order to serve local, regional, and national economies, logistics facilities should be located near regional infrastructure networks. Logistics sprawl can be partially a result of the differences in land prices between suburban–exurban and central urban areas. Moving away from the built-up urban areas is due to land availability and low costs, interconnections with regional and national flows from suburbs, and local policies and governmental interventions. All these factors also explain why the phenomenon of logistics sprawl does not occur in study regions. One earlier study argued that freight activities are more heavily concentrated in central counties in major metropolitan areas in Texas than other US metropolitan areas (Cidell, 2010). Future research could use qualitative research methods to reveal the effects of institutional factors, such as municipal policies, environmental regulations, and financial incentives. As we all know, local government agencies play crucial roles in planning practices and allocating logistics facilities to meet their needs.

Spatial cluster maps between warehouse and e-commerce activities are helpful and informative for planners and policy makers who are interested in reorganizing the spatial distribution of warehouses and other logistics facilities to fulfill customers' needs related to the dynamics of e-commerce supply chains. As e-commerce fulfillment demands spike, warehouse managers and logistics providers need to examine their operational activities for faster, more accurate, and more productive results. Those spatial clusters with more e-commerce establishments but fewer W&DCs also warrant further study on their socioeconomic characteristics, network accessibility, transportation and industrial activities, as well as warehouse space utilization. We do observe that the number of spatial clusters that are worthy of further investigation has been declining in Houston and Dallas–Fort Worth during the study period. There is one major problem with the spatial bivariate analysis presented, namely that it fails to capture the true influence of a particular factor while controlling for other factors.

All statistical tests suggest a much weaker relationship between e-commerce and distribution facilities in Houston than in Dallas–Fort Worth and Austin. The results may be due to the fact that Houston's warehouse activities are heavily reliant on maritime cargo transportation. Internationally, the Port of Houston is connected to many ocean carriers that provide services on all major trade lanes and container shipping routes. With over two-thirds of U.S. gulf coast container traffic handled there, it is the largest gulf coast container port (Greater Houston Partnership, 2021). Accordingly, future studies could conduct multivariate analyses for e-commerce and distribution facilities based on their facility types or employment sizes, pending on the data availability. We also find that changes in e-commerce and manufacturing are more likely to lead to changes in warehousing location choice in the Dallas–Fort Worth regions than in Austin and Houston. Workers of color tend to play a greater role in Austin's warehousing industry. Data-driven machine learning approaches provide nuanced guidance on specific planning efforts because they rank the relative importance of explanatory variables and show nonlinear effects; however, these mechanisms need to be explored in greater depth.

This study focuses e-commerce and warehousing activities in Texas between 2003 and 2016, with a particular focus on three large MSAs, in terms of Dallas–Fort Worth, Austin, and Houston. The research results confirm that, along with e-commerce facilities, transportation activities, transport network accessibility, second-order industrial activities, and other social-economic conditions influence warehousing activities significantly. This study shows that the economic crisis of 2007-2009 influenced warehousing and logistics industry in Texas substantially. Local and regional agencies in Texas play a role in supply chains and logistics development through zoning, tax policies, and industry and workplace standards. Planning and policy makers can use the research results to better understand how logistics and e-commerce travel patterns are associated with urban spatial structure and function. This study also provides useful information to local logistics service providers on determining the appropriate locations of new warehouses and shipping facilities considering transportation planning practices, in collaboration with city and state planning departments. This is particularly important during the post pandemic era as there is much uncertainty about the potential COVID-19 long-term effects. Last but not the least, despite some above indicated limitations, this research proposes a novel framework and uses this framework in the Texas context, which offers important references for warehouse location analysis in the e-commerce era elsewhere.

Acknowledgement

This study was partially supported by the University Of Houston-National Research University Fund (NRUF) (Project Number: R0507403),

References

Alsaleh, N., & Farooq, B. (2021). Interpretable data-driven demand modelling for on-demand transit services. Transportation Research Part A: Policy and Practice, 154, 1-22.

Anselin, L., Syabri, I., & Kho, Y. (2010). GeoDa: an introduction to spatial data analysis. In Handbook of applied spatial analysis (pp. 73-89). Springer, Berlin, Heidelberg.

Bartholdi, J. J., & Hackman, S. T. (2014). Warehouse & distribution science: release 0.96. The Supply Chain and Logistics Institute, 30332.

Boysen, N., De Koster, R., & Weidinger, F. (2019). Warehousing in the e-commerce era: A survey. European Journal of Operational Research, 277(2), 396-411.

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. Travel Behaviour and Society, 14, 1-10.

Cidell, J. (2010). Concentration and decentralization: The new geography of freight distribution in US metropolitan areas. Journal of Transport Geography, 18(3), 363-371.

Dablanc, L., & Rakotonarivo, D. (2010). The impacts of logistics sprawl: How does the location of parcel transport terminals affect the energy efficiency of goods' movements in Paris and what can we do about it? Procedia-Social and Behavioral Sciences, 2(3), 6087-6096.

Dablanc, L., Ogilvie, S., & Goodchild, A. (2014). Logistics sprawl: differential warehousing development patterns in Los Angeles, California, and Seattle, Washington. Transportation Research Record, 2410(1), 105-112.

Demir, E., Huang, Y., Scholts, S., & Van Woensel, T. (2015). A selected review on the negative externalities of the freight transportation: Modeling and pricing. Transportation research part E: Logistics and transportation review, 77, 95-114.

Ding, C., Cao, X. J., & Næss, P. (2018). Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. Transportation Research Part A: Policy and Practice, 110, 107-117.

Dong, W., Cao, X., Wu, X., & Dong, Y. (2019). Examining pedestrian satisfaction in gated and open communities: An integration of gradient boosting decision trees and impact-asymmetry analysis. Landscape and urban planning, 185, 246-257.

Finch, M., Perdue, B., Boske, L. B., & Harrison, R. (2017). Global Logistics Hubs in Texas. Transportation Policy Brief, (3). Link: https://library.ctr.utexas.edu/ctr-publications/5-6690-01/prp3.pdf

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gao, K., Yang, Y., & Qu, X. (2021). Examining nonlinear and interaction effects of multiple determinants on airline travel satisfaction. Transportation Research Part D: Transport and Environment, 97, 102957.

Giuliano, G., & Kang, S. (2017). Spatial Dynamics of Warehousing and Distribution in California. METRANS UTC Draft 15-17 January 2017 (No. CA-17-2640).

Giuliano, G., Kang, S. and Yuan, Q. (2016). Spatial dynamics of the logistics industry and implications for freight flows. A National Center for Sustainable Transportation Summary Report. NCST Project USC-CT-TO-004, METRANS Transportation Center, Sol Price School of Public Policy, University of Southern California, Los Angeles, CA.

Gu, J., Goetschalckx, M., & McGinnis, L. F. (2007). Research on warehouse operation: A comprehensive review. European journal of operational research, 177(1), 1-21.

Guerin, L., Vieira, J. G. V., de Oliveira, R. L. M., de Oliveira, L. K., de Miranda Vieira, H. E., & Dablanc, L. (2021). The geography of warehouses in the Sao Paulo Metropolitan Region and contributing factors to this spatial distribution. Journal of Transport Geography, 91, 102976.

Guerrero, D., Hubert, J. P., Koning, M., & Roelandt, N. (2022). On the spatial scope of warehouse activity: An exploratory study in France. Journal of Transport Geography, 99, 103300.

Guo, J., & Zhang, M. (2021). Exploring the Patterns and Drivers of Urban Expansion in the Texas Triangle Megaregion. Land, 10(11), 1244.

Greater Houston Partnership (2021). Houston Fact, 2021. Retrieved July 5, 2022, from https://www.houston.org/sites/default/files/2021-09/houston%20facts%202021_digital_Final.pdf

Jaller, M., & Pahwa, A. (2020). Evaluating the environmental impacts of online shopping: A behavioral and transportation approach. Transportation Research Part D: Transport and Environment, 80, 102223.

Jaller, M., Pineda, L., & Phong, D. (2017). Spatial analysis of warehouses and distribution centers in Southern California. Transportation Research Record, 2610(1), 44-53.

Jaller, M., Qian, X., & Zhang, X. (2020). E-commerce, Warehousing and Distribution Facilities in California: A Dynamic Landscape and the Impacts on Disadvantaged Communities.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jin, T., Cheng, L., Liu, Z., Cao, J., Huang, H., & Witlox, F. (2022). Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments. Transport Policy, 117, 48-59.

Kang, S. (2020a). Warehouse location choice: A case study in Los Angeles, CA. Journal of Transport Geography, 88, 102297.

Kang, S. (2020b). Why do warehouses decentralize more in certain metropolitan areas?. Journal of Transport Geography, 88, 102330.

Koushik, A. N., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behaviour research: a review. Transport Reviews, 40(3), 288-311.

Mahoney, N., 2020, October 12. As e-commerce soars, logistics real estate in Texas is hot. FreightWaves. Retrieved March 22, 2022, from https://www.freightwaves.com/news/as-online-shopping-soars-logistics-centers-are-booming-in-texas.

Mao, T., Mihăită, A. S., Chen, F., & Vu, H. L. (2021). Boosted genetic algorithm using machine learning for traffic control optimization. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7112-7141.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

Onstein, A. T., Tavasszy, L. A., & van Damme, D. A. (2019). Factors determining distribution structure decisions in logistics: a literature review and research agenda. Transport Reviews, 39(2), 243-260.

Rai, H. B., Kang, S., Sakai, T., Tejada, C., Yuan, Q. J., Conway, A., & Dablanc, L. (2022). 'Proximity logistics': Characterizing the development of logistics facilities in dense, mixed-use urban areas around the world. Transportation Research Part A: Policy and Practice, 166, 41-61.

Sakai, T., Beziat, A., & Heitz, A. (2020). Location factors for logistics facilities: Location choice modeling considering activity categories. Journal of Transport Geography, 85, 102710.

Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. Accident Analysis & Prevention, 122, 226-238.

Tao, T., Wang, J., & Cao, X. (2020). Exploring the non-linear associations between spatial attributes and walking distance to transit. Journal of Transport Geography, 82, 102560.

Thomas, L. (2021, February 24). Gap to invest $140 million to build Texas warehouse as online sales swell. CNBC. Retrieved March 22, 2022, from https://www.cnbc.com/2021/02/24/gap-to-invest-140-million-into-texas-warehouse-as-online-sales-swell.html

US Department of Commerce. (2021a). Retail e-commerce sales in the United States from 1st quarter 2009 to 3rd quarter 2021. In Statista - The Statistics Portal. Retrieved February 16, 2022, from https://www.statista.com/statistics/187443/quarterly-e-commerce-sales-in-the-the-us/

US Department of Commerce. (2021b). E-commerce as share of total US retail sales from 1st quarter 2010 to 3rd quarter 2021. In Statista - The Statistics Portal. Retrieved February 16, 2022, from https://www.statista.com/statistics/187439/share-of-e-commerce-sales-in-total-us-retail-sales-in-2010/

Wang, K., & Ozbilen, B. (2020). Synergistic and threshold effects of telework and residential location choice on travel time allocation. Sustainable cities and society, 63, 102468.

Woudsma, C., & Jakubicek, P. (2020). Logistics land use patterns in metropolitan Canada. Journal of Transport Geography, 88, 102381.

Woudsma, C., Jakubicek, P., & Dablanc, L. (2016). Logistics sprawl in North America: methodological issues and a case study in Toronto. Transportation Research Procedia, 12, 474-488.

Xiao, Z., Yuan, Q., Sun, Y., & Sun, X. (2021). New paradigm of logistics space reorganization: E-commerce, land use, and supply chain management. Transportation Research Interdisciplinary Perspectives, 9, 100300.

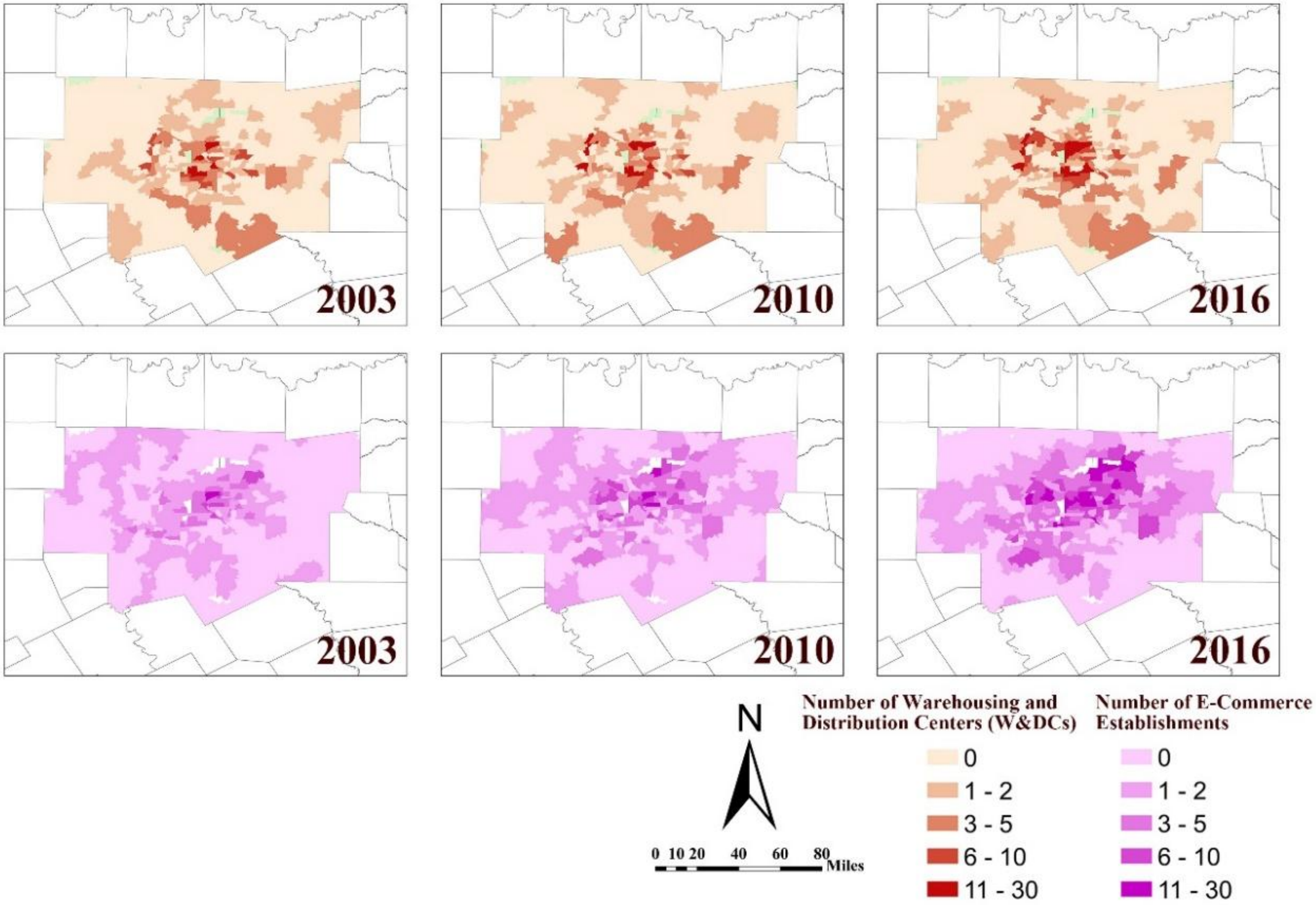
Yeates, M. (1974): An Introduction to Quantitative Analysis in Human Geography, McGraw Hill, New York

Yang, H., Zhang, Q., Helbich, M., Lu, Y., He, D., Ettema, D., & Chen, L. (2022). Examining non-linear associations between built environments around workplace and adults' walking behaviour in Shanghai, China. Transportation Research Part A: Policy and Practice, 155, 234-246.

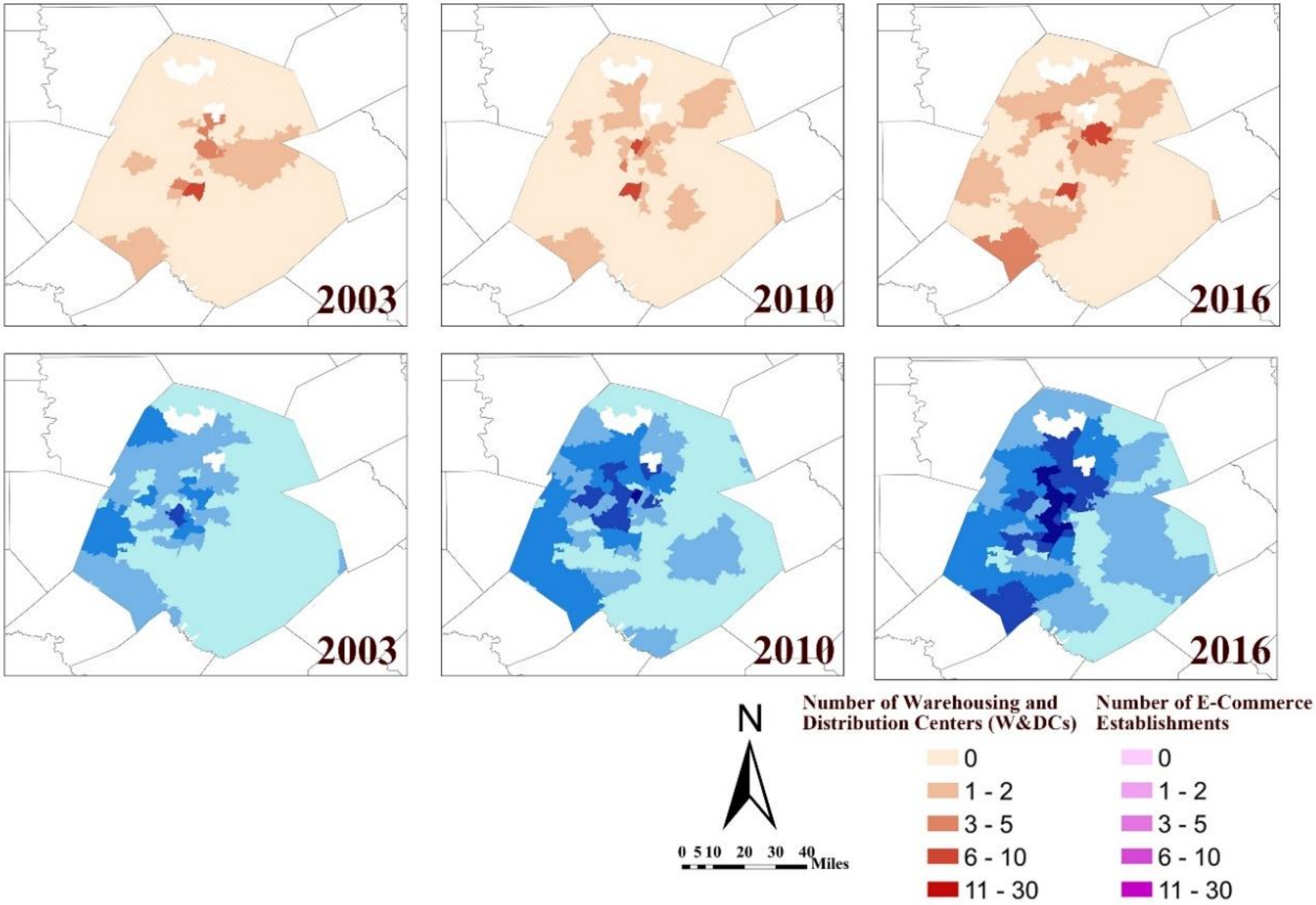
Yuan, Q. (2019). Does context matter in environmental justice patterns? Evidence on warehousing location from four metro areas in California. Land use policy, 82, 328-338.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324.

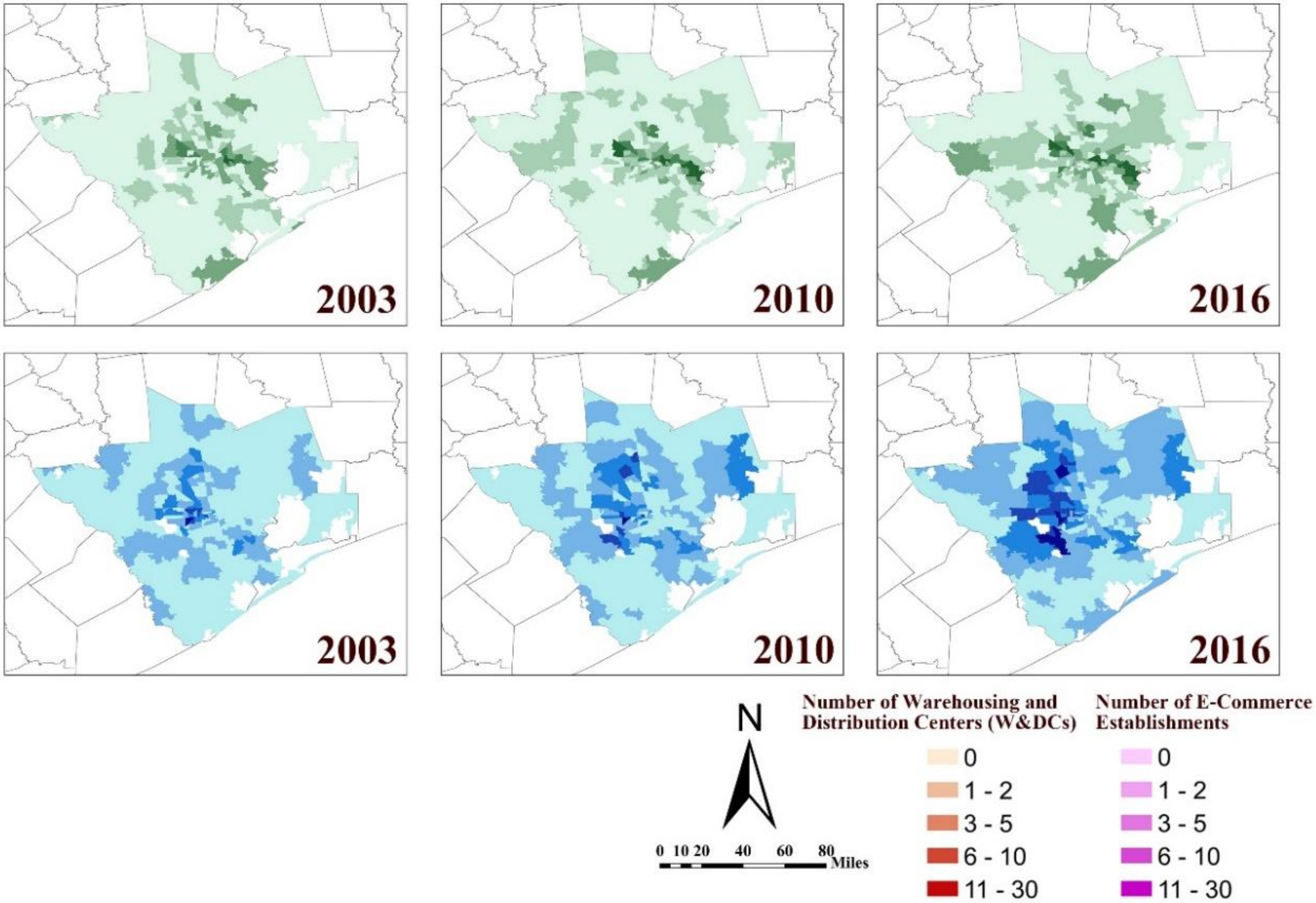
Appendix A



(a) Number and location of W&DCs and e-commerce establishments in Dallas–Fort Worth



(b) Number and location of W&DCs and e-commerce establishments in Austin



(c) Number and location of W&DCs and e-commerce establishments in Houston

Figure A1. Number and location of W&DCs and e-commerce establishments in three MSAs

Table A1. OLS Regressions

	All		Dallas		Austin		Houston		Not in three MSA	
	Coeff.	VIF	Coeff.	VIF	Coeff.	VIF	Coeff.	VIF	Coeff.	VIF
E-Commerce Facilities										
Number of e-commerce establishments	0.238***	1.50	0.293**	1.70	0.143***	1.79	0.153***	1.65	0.174**	1.31
Transportation Activities										
Number of air transport establishments	0.061	1.07	0.076	1.10	0.014	1.21	0.065	1.10	0.477*	1.17
Number of water transport establishments	0.448**	1.16	2.107*	1.09	-0.477**	1.03	0.295*	1.19	0.872	1.04
Access to Transportation Infrastructure										
Distance to highway	-0.074***	1.37	-0.029	1.50	-0.005	1.47	-0.121***	1.52	-0.005	1.34
Distance to seaport	0.003***	1.32	0.000	1.22	0.005	3.60	-0.009	2.22	0.000	1.20
Distance to airport	0.003	1.38	-0.058	1.37	-0.021	1.48	0.086**	1.59	0.010	1.32
Distance to intermodal facility	-0.010***	1.39	-0.017	3.54	-0.006	2.93	-0.028*	3.79	-0.002*	1.60
Industrial Activities										
Manufacturing intensity	0.191***	1.66	0.187***	1.89	0.265*	4.03	0.158**	1.90	0.175**	2.32
Retail trade intensity	-0.024***	2.67	-0.033**	2.95	-0.059*	6.48	-0.016*	2.72	-0.016*	2.51
Population Density	-0.232***	2.94	-0.286***	3.15	0.091	5.06	-0.293***	3.58	-0.009	2.34
Median Household Income (in $1000)	-0.007*	2.70	-0.007	3.06	-0.010**	3.45	-0.004	2.96	-0.006*	1.85
Race (base case: % of population is white)										
% of population is Black	0.780	1.46	0.100	1.95	-1.118	2.08	0.545	1.49	1.159***	1.31
% of population is Asian	3.246*	1.81	3.029	1.80	1.893	2.38	3.245	2.07	-3.089	1.37
% of population is other races	-0.777	1.62	-3.138	1.69	0.769	3.27	3.142	1.65	-0.260	1.21
Ethnicity (base case: % of population is non-Hispanic)										
% of population is Hispanic	3.468***	2.93	5.150***	3.32	1.078	5.18	2.310**	3.09	1.235**	1.97
Housing occupancy status (base case: % of occupied housing units)										
% of vacant housing units	1.134	1.80	0.496	2.14	-0.608	2.37	1.664	2.08	-1.081**	1.46
Internet subscriptions in household (base case: % of households do not have internet access)										
% of households have internet subscription	0.698	2.80	-0.105	3.40	-0.076	4.47	0.474	2.92	1.422*	2.03
% of households have internet access without a subscription	-0.129	1.27	-1.538	1.39	-1.508	1.85	-1.306	1.22	4.745*	1.22
Year										
2004	0.101***		0.086*		0.061		0.115**		0.021*	
2005	0.089***		0.078		0.069		0.081		0.028	
2006	0.136***		0.130*		0.117*		0.122**		0.058***	

2007	0.168***	0.147*	0.189**	0.159*	0.073***
2008	0.075	0.075	0.020	0.112	0.100***
2009	0.071	0.063	0.002	0.124	0.115***
2010	0.094	0.068	-0.039	0.215**	0.108***
2011	0.011	-0.063	-0.075	0.150	0.133***
2012	-0.025	-0.085	-0.201*	0.176	0.098***
2013	0.027	-0.037	-0.127	0.228**	0.131***
2014	0.012	-0.051	-0.123	0.203*	0.123***
2015	0.036	-0.019	-0.108	0.242*	0.146***
2016	0.037	-0.022	-0.168	0.292**	0.162***
Constant	-0.734	1.135	0.691	-0.531	-0.959
R-squared	0.265	0.337	0.293	0.243	0.166
Number of obs.	8,554	4,004	1,316	3,234	13,986

Note: ***p value < 0.01, **p value < 0.05, *p value < 0. 1.