RERI manuscript No.
(will be inserted by the editor)

Dis-Aggregated Urban Location and Commercial Real Estate Values

Jeremy Gabe, University of San Diego
Andy Krause, Zillow
Spenser Robinson, Central Michigan University
Andrew Sanderford, University of Virginia*

May 2021: RERI Draft Version

Abstract This paper examines the relationship between location, specifically urban spatial structure, and commercial real estate capitalization rates. Here, a novel control vector describes non-linear bid-rent curves, enabling exogenous spatial control in commercial real estate pricing models with non-random spatial observation or treatments, such as commercial transaction data. First, the paper estimates this new control vector, a series of continuous form urban spatial structure factors, at the Census Block Group level using Principal Components Analysis, which offer a simplified and alternative methodological pathway to similar econometric techniques such location grids, spatial fixed effects, and spatial autocorrelation. Then, these exogenous measures of location are evaluated for utility in hedonic pricing models where transaction data or treatments are scarce or spatially biased, such as the thinness of data on office and retail transaction cap rates. Results indicate that across each of the asset classes, transaction prices and capitalization rates are both statistically and economically sensitive to variation in the spatial factors. In addition, spatial models of risk demonstrate that retail cap rates are more sensitive to location than office cap rates. Practically, spatial models of risk can improve the allocation of real estate investment capital in complex polycentric urban markets. Finally, the utility of this new exogenous spatial control vector is tested alongside traditional spatial valuation techniques in a forecasting context: automated valuation models of the single-family housing market, where transaction data is not scarce. This spatial control vector produces similar accuracy statistics as traditional approaches in forecasting.

JEL: R20, R31, R40

Keywords Urban Form · New Urbanism · Walkability · Multi-family Housing · Density · Design · Destination Accessibility · Transit

Introduction

The understanding of the relationships between location, land use, and property values continues to evolve across urban economics, real estate, and finance (Alonso, 1968; Anas et al., 1998; Saiz, 2010). The debate about these relationships helps to enhance urban economic, asset pricing, default, transportation and other related models (An and Pivo, 2020; Kok et al., 2014; Titman

Address(es) of author(s) should be given

et al., 1985; Bialkowski et al., 2021). This paper is motivated by two problems common to real estate analysis incorporating this debate.

The first problem is the limited consensus about how to measure the spatial characteristics of non-linear bid-rent curves, particularly outside traditional monocentric models (i.e. in suburbs and polycentric cities), that shape urban land price variation (Wheaton, 1979; Wieand, 1987; Anas and Kim, 1996). The second is the potential for location value and price effects to be attributed to and confounded by exogenous factors that are not spatially independent (Bourassa et al., 2020). For example, the growing popularity of research on estimated price effects of real estate treatment attributes (i.e., green building certification) can be confounded by spatial bias of treated assets (Fuerst and McAllister, 2011). These are significant problems for institutional investors with spatially diversified portfolios seeking to identify new investment opportunities.

The emergence of "big data" (Glaeser et al., 2018b; Bourassa et al., 2020) provides a potential innovative solution with recent work in the urban economics, planning, real estate, and transportation literature examining the role of direct and alternative spatial measurement approaches (Ewing and Cervero, 2010; Kuang, 2017; Bourassa et al., 2020; Gabe et al., 2021; Fisher et al., 2020). However, there are limits to the value of big urban data (Glaeser et al., 2018b); not all of it adds value or helps clarify the complexity within urban economic spatial relationships. These limits create the opportunity for this paper; the opportunity to generate simplified large data measures of location that capture spatial and economic relationships.

This paper explores whether "big data" can be used in a variance reduction framework to extract latent factors of urban spatial structure. For example, some measure of density (i.e. employment) often is used to proxy land use intensity, though it is often noisy at small spatial scales. Here, this paper uses Principal Component Analysis (PCA) to generate two latent and non *a priori* factors from the Environmental Protection Agency Smart Location Database (SLD), over 90 measured characteristics of urban spatial structure.

The resulting latent "Factor 1" describes urban economic activity intensity across entire polycentric urban forms as a continuous variable constructed from urban form attributes traditionally associated with a central business district (CBD). Latent "Factor 2" is orthogonal, a feature of PCA, and identifies lower density suburban or exurban spatial structure akin to Hoyt's concept of the outer ring. Combined, these factors represent a non-linear continuous measurements of location at the Census Block Group (CBG) level. Contributing inputs to the two latent factors, and thus their subjective interpretation, are consistent across major U.S. Metropolitan Statistical Areas.

As a result, these latent factors offer a potential new methodological pathway to complement other econometric and geospatial controls for urban spatial structure such as location (x,y) grids, spatial fixed effects, and spatial autocorrelation. Importantly, an exogenously measured control for location could evaluate, using a non-linear bid-rent curve, the multi-dimensional economic relationships between location and prices or rents, which largely consists subjective definitions of spatial sub-markets or arbitrary identification of central business districts.

A further advantage of the latent factors of urban spatial structure explored here is that they are measured exogenously to a dataset of observed prices or transactions. Existing techniques to control for location in asset pricing models - spatial fixed effects, (x,y) spatial grids, or spatial autocorrelation - rely on econometric techniques to infer the value of location from the observed data. In contrast, an exogenous measure of location could control for location value in data samples lacking statistical power to endogenously control for spatial effects, such as spatially variant samples with relatively low numbers of observations.

Commercial real estate transaction models are a prime application for this technique. Office and retail asset sales are relatively thin in a given market when compared with housing sales, so current modelling strategies must add a lot of spatial and/or temporal variance (large areas

in the sampling frame and/or with multiple years of observations). To examine the potential of latent Factors 1 and 2 to address the problems and opportunities identified above, both Factors are incorporated into hedonic regression models analyzing a 10-year sample of office and retail transaction prices and capitalization rates from Real Capital Analytics. The sample contains approximately 40,000 office and retail transaction observations in more than 35 U.S. Core Based Statistical Areas (CBSA). Approximately 25% of these transactions have income data to evaluate cap rates. One immediate practical goal is to examine, and map, the relationship between these exogenous measures of location and cap rates, enabling the capital markets to evaluate the risk premium placed on any Census Block Group in a CBSA of interest.

To explore the robustness of this spatial control, this research evaluates how well these latent spatial factors perform compared to traditional spatial controls, a question similar to that of Bourassa et al. (2020), but with a different proposed measure of urban activity. Out of sample testing is conducted on single-family detached housing data, where data sample sizes are traditionally seen as sufficient for endogenous spatial controls.

The paper addresses three research questions: 1) In both office and retail, where spatial heterogeneity is greater than in single-family housing, how if at all, are the new latent measures of urban spatial structure related to commercial real estate pricing and risk evaluation (cap rates)? 2) To what extent, if any, can latent factors of urban spatial structure identify and map spatial relationships associated with commercial real estate pricing and risk? And 3) to what extent, if any, do the latent factors suggest capacity to attenuate issues of spatial endogeneity in real estate pricing models?

The results suggest two contributions to the literature. First is a proof of concept. Models indicate that across each of the asset classes, both transaction prices and capitalization rates are both statistically and economically sensitive to variation in the continuous form specifications of each PCA factor. This leads to practical insights, notably that retail risk is more location-sensitive than office, which has less spatial heterogeneity. Accuracy of predictions using PCA Factors in housing price models show similarities with current techniques of spatial control. This suggests that PCA Factors 1 and 2 produce useful exogenous location signals at small spatial resolution, and can be evaluated for any Census Block Group in the United States. This method is statistically efficient at low sample sizes; the smaller sample size for cap rates (n 10,000 over 10 years and 35 markets) relative to the sample for transaction pricing (n 40,000 over 10 years and 35 markets) produce nearly identical maps of spatial market variance.

Second, the use of latent PCA factors of urban spatial structure "big data" can control for spatial bias in real estate models similar to current methods but the exogenous nature of these latent factors fills a niche to control for spatial bias in models with small sample sizes. Factors 1 and 2 add statistically significant though only marginal economic value to out-of-sample forecasting models using traditional spatial approaches (e.g., small-scale fixed-effects or spatial autocorrelation when sample sizes are large). This finding is congruent with evidence from analyses by Bourassa et al. (2020) suggesting that big data might not provide a panacea and that indirect or endogenous measurements may, in appropriate circumstances, offer an equally acceptable methodological approach in regards to forecasting accuracy. However, it is also clear that the the PCA factors do help isolate and control for spatial bias when sample sizes are too small for traditional spatial modelling approaches, a particular problem when treatment attributes do not have a random spatial distribution.

Background & Identification Framework

Urban Economic Models & Property Prices

The traditional urban model provides an explanation of relationships between urban form and property prices. It originally identified an inverse linear relationship between central locations and land rents or property prices (Alonso, 1968); this relationship is a bid-rent curve describing the decrease in willingness to pay for proximity to centrality and the attendant increase in commuting costs Muth (1975). In its simplest application, the model assumes a monocentric urban area on a flat plane, where housing can occur in any location and location value is based on access to the monocentric urban center.

However, urban areas are rarely monocentric and are often constrained by idiosyncratic geography (Saiz, 2010). Further, households and firms with differing levels of wealth also attempt to satisfy multi-faceted utility functions that incorporate public service quality, dis-amenity and tax avoidance, and other factors besides accessibility (Wheaton, 1974). They may not select locations with the greatest proximity to traditional central amenities in favor of satisficing or other decision-making paradigms. There is, as a result, substantial debate about how to easily measure the spatial and economic drivers of urban property prices (Anas and Kim, 1996; Ahlfeldt and Wendland, 2013; Fisher et al., 2020).

Competing explanations draw on agglomeration spillover models where complementarities and dynamics between firms (and labor) sharing public good inputs explain urban land price variation (Eberts and McMillen, 1999). Other models postulate a dynamic urban spatial structure equilibrium predicated on the ways that firms balance the exogenous benefits from co-locating near other producers against their workers commuting costs (Lucas and Rossi-Hansberg, 2002). Assumed in these alternative explanations, is the notion of poly-centricity and that housing or commercial activity can occur in a multiplicity of spatial formats.

These different and competing perspectives illustrate the complexity of urban land markets where location supply inelasticity clashes with differing tastes and preferences of households and firms (Roback, 1982; Titman et al., 1985). They also demonstrate that variation in property prices is materially related to multiple dimensions of location such as the diversity of land use, density, and dimensions of accessibility to and across the urbanized landscape (Glaeser et al., 2018b). They also reveal the limited consensus about how to easily measure the spatial characteristics of non-linear bid-rent curves, one of the problems motivating this paper. Converting this complexity into a functional form for pricing models is a major topic of real estate research.

0.1 Modelling Cap Rates and Commercial Real Estate Investment Risk

In commercial real estate, there is a literature examining the factors that contribute to variation in cap rates—including analyses of location (Peng, 2016; Bialkowski et al., 2021). Of note and significance here is work by Fisher et al. (2020) analyzing the relationship between location and performance of public REITs. They contend that supply inelasticity in urban spaces creates advantages over those in suburban locations; that rent and price changes should be greater in supply constrained locations and that those locations are likely to have greater capacity to withstand supply shocks.

Urban Spatial Structure Data

Innovation in empirical information on urban spatial structure has evolved with advances in remote sensing, geographic information systems, and detailed longitudinal surveys (Naik et al., 2016). Emerging "big data" adds potential to enhance the descriptoin of urban spatial structure rather than more oblique proxy measures in common use, like distance to a node of interest (Chetty et al., 2014; Glaeser et al., 2018b). Early work using such data in real estate modelling involved "walkability" indices such as WalkScore, which provides a composite index describing the walkability, or friendliness of the urban form to walking for recreation or transportation, of a particular location. Evidence indicates that walkability is positively associated with office and multi-family property prices (Pivo and Fisher, 2011; Bond and Devine, 2016)). More recent measurement advances include the use of consumer review or mobile phone tracking data to describe travel behavior at different scale and frequency than traditional survey measures. The evidence suggests that in some cases, such "big data" can be additive to analyses of urban phenomena (Glaeser et al., 2018a; Kuang, 2017). In other cases, despite the novelty, big data fails to add value to existing modelling techniques (Bourassa et al., 2020).

Specific to real estate pricing models, small scale geographic fixed effects appear to better instrument for the spatial value of a particular location in housing markets than novel mobile phone tracking data used to describe transport patterns (Bourassa et al., 2020). But not all real estate pricing models can use small scale fixed location effects, which are available in large or spatially concentrated sample sizes like single-family housing sales in urban centers.

This insight raises the possibility of differentiated outcomes in the utility of "big data" applied in the description of urban spatial structure. Could "big data" have utility in questions that can only be addressed with sample-size constrained data or at large spatial scales? Currently, basic exogenous measures such as distance to a pre-determined central business district (CBD) or institutional definitions of sub-markets are used in models such constrained contexts and have limitations that novel geospatial data may overcome. As an example of these limitations on current practice, brokerage houses and other real estate market participants have modified institutional definitions, so sub-market definitions can be dynamic and vary depending on the brokerage house. In the case of CBDs, the Census defined boundaries align with a specific Census Tract (CT) or multiple Tracts (Limehouse and McCormick, 2011), but institutional sub-market boundaries may or may not align with Census geographies. That leads to problems as many demographic and other relevant variables of interest to real estate market modellers are produced using Census geographies. As a result, there appears to be an untested opportunity for novel "big data" urban spatial structure metrics to increase the utility of sample-constrained real estate pricing or risk models.

Recognizing this opportunity, urban geography research suggests potential for a novel spatial location index based on exogenous measurements of urban form and transportation infrastructure that go beyond walkability. This work assumes efficient transportation infrastructure allows less ideal locations to substitute for those highest in demand by reducing the slope of the bid-rent curve (Alonso, 1968). Further, it relies on the notion that urban residents seek to maximize spatial utility, but are forced into trade-off decisions by budgets and competing preferences to centrality of location. Centrality has many aspects, including (among others): *reach* (Sevtsuk and Ratti, 2010), a measure of how many places are directly accessible from a specific location within a restricted distance; *gravity* (Hansen, 1959), the degree to which directly accessible areas are nearby or far away from a desired location, such as a CBD; *betweenness* (Freeman, 1977), the frequency that a specific location is on the shortest path between any two areas; *closeness*(Sabidussi, 1966), or how central a specific location is to all other places in the urban area; and *straightness* (Porta et al., 2009), or whether the transportation network is designed

such that distances from a specific location to other places match with Euclidean (straight-line) distances.

In 2015, the EPA Smart Location Database (SLD) was first produced to gather innovative empirical data on urban form, which can be used to identify and describe location efficiency across multiple dimensions of urban form (Song and Knaap, 2004; Ewing and Cervero, 2010). The SLD provides direct measurement of urban form across five distinct categories, or dimensions: density, land use diversity, design, destination accessibility, and distance to transit. Within each dimensions, the SLD provides a number of detailed individual metrics (e.g., employment density across 10 different industrial classifications) in continuous form specification (these data are described in further detail below). Measuring urban spatial structure directly and across dozens of metrics, it offers an opportunity to use "big data" to improve exogenous and non-linear measurement of urban spatial structure, which was previously latent, elliptical, or formed by *a priori* assumptions (Glaeser et al., 2018b).

Integrated Framework and Expectations

The identification framework proposed to exploit the SLD in real estate modeling, while recognizing competing explanations for the relationships between property prices and urban form, the differences and challenges in measurement techniques underpinning them, and the potential for location value and price effects to be attributed to and confounded by exogenous factors that are not spatially independent.

Differing from recent approaches in Fisher et al. (2020), the framework assumes that each unit of geography (e.g., a neighborhood or Census Block Group) can have multiple concurrent spatial ontologies and be formed by multiple spatial economic factors. It assumes that these dimensions include both urban and the suburban—attributes that can be endogenous. Following Glaeser et al. (2018b), the framework expects that the SLD increases the precision of measuring urban spatial economic concepts including land use diversity, building and employment density, urban design, accessibility, and proximity to transit amenities. Moreover, it assumes that several continuous form measures can be derived from the SLD data using variance reduction strategies to extract latent signals, a *non a-priori* framework, and that the resultant continuous form latent signals have the potential to measure the complex non-linear bid-rent curves for that unit of geography. In other words, rather than using a single measure or estimating a single bid rent curve for an entire market, even a poly-centric one (Heikkila et al., 1989)), the new data makes it possible to estimate continuous bid-rent curves for a geographic unit that can be updated on future releases of the SLD.

With respect to commercial real estate, the framework expects that these continuous latent urban form measures will be statistically and economically significant predictors of office and retail transaction prices and capitalization rates (McMillen and McDonald, 1997; Pivo and Fisher, 2010; Fisher et al., 2020). It also assumes that prices and risk respond to both urban and suburban characteristics (Bourassa et al., 2007), and that a particular urban location can have overlap of urban and suburban structure. Such factors should present with opposite effects; for example, urban oriented factors are expected to present positively in office transaction prices (and negatively in cap rates) while the suburban oriented factors are expected to present oppositely given the spatial distributions of the asset classes (Fisher et al., 2020). Measuring "urban" and "suburban" separately allows a non-linear relationship between price (or risk) and location efficiency.

Data

The data used here combines two types of information: 1) building level observations describing both structural and economic attributes for a sample of office and retail buildings within 35 Core Based Statistical Areas in the United States and 2) an array of urban spatial structure data from the Smart Location Database as "big data" with potential to measure spatial structure exogenously. All building level data is from the Real Capital Analytics Transaction Database. Additional control variables come from the U.S. Census American Community Survey 5-year Estimates, US Census "Tiger" shapefiles, and Applied Geographic Resources. The common spatial scale for this analysis is the US Census Block Group (CBG).

There are two analysis data sets. The first is the EPA SLD. It is used to estimate the latent urban form factors used in later models. The second analysis data set is constructed by matching each property (by geographical coordinates) with its relevant Census Block Group. This enables the integration of the urban spatial structure data using Federal Identification Processing Standard (FIPS) codes unique to each Block Group. For the extraction of latent variables of urban spatial structure, the Block Group is the unit of analysis. For pricing models, a unique commercial property is the unit of analysis; each row of information contains cap rate and/or transaction observations for a building, physical descriptors of that property, the latent urban form variables generated from the spatial structure data (detailed below), and auxiliary location information for the CBG. (Gordon-Larsen et al., 2006; Song and Knaap, 2004).

Smart Location Data

The EPA Smart Location Database (SLD) provides measures of several demographic, employment, and built environment variables for every CBG in the United States (Ramsey and Bell, 2014). The SLD contains more than 100 measures of across five dimensions of urban form: land use density, land use diversity, urban design, destination accessibility, and transit proximity. Index variables included in the SLD package that are derived from individual SLD measures, such as the EPA Walkability Index, are excluded. The SLD contains a number of measures from the General Transit Feed Specification (GTFS). However, the GTFS data does not exist for the entirety of the 35 markets in the sampling frame. Consequently, all GTFS measures are excluded. This reduces the SLD data to 90 unique measures of urban form available for nearly all CBGs in the United States.[1]

To describe the dimensions of urban form that produce these 90 input variables, *Density* measures housing units per acre, population per acre, and jobs per acre by industrial classification. *Land Use Diversity* describes the different land uses within an area. Specific factors measuring land-use diversity include jobs-housing balance, employment entropy and trip generating estimates based on employment diversities. *Design* describes elements of physical and transport infrastructure and the bias of each type of design relative to its users (e.g., cars, transit, or people). Individual metrics detail the total road network density, network density for various use modalities and intersection density by intersection type. *Distance to Transportation* summarizes access and quality of nearby fixed guideway public transport. Fixed guideways describe rail and bus infrastructure with exclusive rights-of-way. Specific measures include the proportion of the CBG employment within one-quarter and one-half mile buffers of transit stops and the frequency of transit service within a CBG. *Destination Accessibility* describes proximity and accessibility to and across the city by various of modes of transportation. Individual metrics of destination

[1] CBGs in the state of Massachusetts do not include 10 employment data variables in the SLD, so Boston and other CBSAs that include Massachusetts CBGs are run on 80 variables.

accessibility measure a range of accessibility indicators derived from engineering and structural equation models drawing on transit patterns, trip generation matrices, employment, and housing patterns.

Variable definitions for each measure in the SLD are provided in the Appendix.

Property Information

Building level observations are drawn from the Real Capital Analytics Property database. This database captures information about building and structural attributes (e.g., ownership, square feet, age, etc). It also includes economic and transactional details (e.g., recent transaction prices, transaction dates, capitalization rates). RCA provides location details for each building including latitude/longitude; for the office product type, it identifies whether or not the property resides within a pre-determined Central Business District of its respective market. Relevant property level variables used in the analysis are summarized in Table 1.

Office building data reflect mean prices in CBD areas exceeding $85 million and suburban office near $18 million. These price points reflect the institutional character of the data set, although there is some right skew in the CBD cohort from a small number of large purchases. Malls and strip malls have mean prices of roughly $12 million and $11 million respectively.

Modeling and Identification Strategies

In the context of the aforementioned research opportunity for a novel exogenous measure of location efficiency for use in real estate models, a two-step method first identifies the new metric and then tests its utility in a variety of real estate modelling contexts. This first step uses Principal Component Analysis (PCA) as a variance reduction strategy to generate a set of latent urban form factors derived from the SLD. The second step evaluates consistent latent factors in a traditional hedonic framework to analyze the relationships between urban spatial structure and commercial real estate prices and yields (Rosen, 1974; Sivitanides et al., 2003), specifically in the context of commercial real estate risk models, where observations are thin and spatially diverse. Out of sample testing in single-family housing is also employed to assess construct validity relative to existing strategies of location control.

Principal Component Analysis of the Smart Location Database - Methodology

Principal Component Analysis (PCA), as a variance reduction method to extract latent signals out of noisy data has been used in the real estate and urban economics literature for measuring real estate returns (Cotter and Roll, 2015), sentiment (Heinig and Nanda, 2018), and urban vibrancy (Barreca et al., 2020). The detailed matrix algebra behind PCA is detailed in Ringnér (2008); Wold et al. (1987); Abdi and Williams (2010) and briefly summarized below.

The innovation tested here is that the 90 variables in the SLD are noisy proxies for latent measures of urban spatial structure, and PCA can extract these latent signals to produce continuous measures of location relevant to real estate markets. For example, auto ownership is expected to be higher in suburban spaces and lower in urban spaces and employment densities are concentrated in CBDs and other commercial centers. PCA uses this noisy data to find orthogonal vectors that describe continuous measures of "urban" and "suburban" spatial structure (and other urban form characteristics) using the variance observed across the entire dataset. These

vectors, which can be described qualitatively by interpreting correlations with the input data, are unobservable latent measures of urban spatial structure derived from what can be observed.

To generate these latent vectors, consider the SLD as N × X matrix S, where N is the number of Census Block Groups (i.e. observations) and X the number of SLD input variables (90). As each variable in X has different units, scale each variable to unit variance to allow for equal weighting by variable. Next, compute eigenvalues (λ) of the covariance matrix of S (KXN):

$$det(K_{XN}) - \lambda I = 0 \tag{1}$$

Where I is the identity matrix matching the dimensions of the covariance matrix (X × X). There will be X eigenvalues. For each eigenvalue, calculate the corresponding eigenvector of length X. Arrange these eigenvectors in order based on their eigenvalues from highest to lowest to arrive at X × X matrix, R, which preserves all variance observed in KXN (the covariance matrix of S).

A key characteristic of R is that the eigenvalues, in descending order, represent the amount of variance communicated by each eigenvector; practically these are interpreted as the latent factors underlying the observed survey data. Removing the eigenvectors with the smallest eigenvalues (e.g. those on the right-side of R) has minimal effect on the overall variance remaining in matrix R, an efficient means of reducing dimensionality. Furthermore, as a characteristic of eigenvectors, each remaining eigenvector is uncorrelated with all preceding eigenvectors.

The decision on how many eigenvectors to retain is inherently subjective. Eigenvalues collectively sum to the total variance observed. With each variable having unit variance, the eigenvalues of S will sum to X (i.e. the number of variables in the SLD). Thus, one approach is to discard any eigenvalue less than 1, as it contains less information than an original input variable. Another approach is to start from the left and cumulatively sum eigenvalues until reaching a certain percentage of the original variance; 50%, 60%, and 80% are all common thresholds.

A third approach, used here, is to concentrate on the practical interpretation of each eigenvector, which is less rigid, but more meaningful than the above approaches. In essence, start with the eigenvector associated with the highest eigenvalue, then move to the next highest. To "interpret" each eigenvector in practical terms, the authors evaluate the 10 input variables which have the 10 highest positive covariances with the subject eigenvector and the variables with the 10 highest negative covariances (i.e. inverse relationships with the subject eigenvector). These covariates premise an argument describing the nature of the subject eigenvector in practical terms.

To further support to the identifying conclusion in this uniquely spatial survey context, we calculate a "factor score" for each observation N by multiplying S by the transposed vector k(λ), where k=1,. . . ,K, or the specific eigenvector associated with eigenvalue λ. The resulting factor score vector can be mapped spatially since each observation is a Census Block Group and compared with the identification conclusion by covariate assessment. If the authors are comfortable identifying the latent contribution represented by the eigenvector with the highest eigenvalue, the identification process moves on to the next highest eigenvalue and so on until covariates and spatial maps are unable to support a clear identifying conclusion.

The factor scores for each Census Block Group, derived from all eigenvectors retained in R are proposed exogenous measures of urban spatial structure. As is typical lexicology in PCA, each retained eigenvector is referred to as a *Principal Component* or *Factor*, numbered from highest eigenvalue to lowest eigenvalue. In this analysis *Factor 1* is the eigenvector associated with the largest eigenvalue.

Finally, PCA on the SLD was first run at the national scale, with all 220,653 U.S. CBGs in one variance reduction model. In addition an independent PCA was evaluated using solely the Census Block Groups in each U.S. Core Based Statistical Area (CBSA) - a common Census designation

used for a metropolitan area. This latter strategy results in different covariance between each latent factor and the underlying SLD metrics for that CBSA. Importantly, CBSA-scale PCA analysis results in potentially different subjective interpretations of each latent factor in each CBSA; for example, "Factor 5" in the Atlanta CBSA PCA is not necessarily a comparable latent measure of urban spatial structure as "Factor 5" in the Seattle CBSA PCA.

0.2 Principal Component Analysis of the Smart Location Database - Results

After performing the national-scale and CBSA-scale PCA on the SLD, a noticable pattern emerged in the interpretation of the latent factors of urban spatial structure. No matter the scale, "Factor 1" can always be interpreted as a continuous variable measuring hubs of commercial activity; strong covariates were predominantly in the density dimension of the SLD. "Factor 2" could always be interpreted as a continuous variable measuring exurban or urban fringe location; strong covariates were always automobile ownership and the land use diversity dimension. Figure 1 describes the raw SLD variables with the largest contribution to Factor 1 and Factor 2 in the national-scale PCA of all 90 SLD variables.

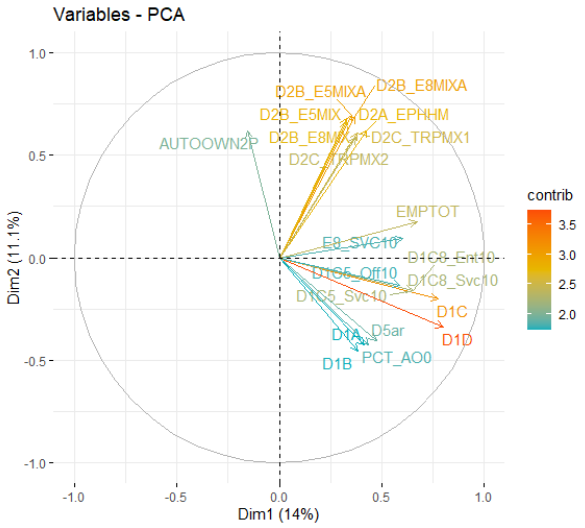

Fig. 1: Visualisation of the 20 strongest contributing variables to Factor 1 and Factor 2 (Dim1 and Dim2 respectively) in the national-scale PCA. Results are nearly identical for every CBSA-scale PCA. Variable definitions described at Ramsey and Bell (2014).

Beyond Factor 2, interpretation of additional latent variables occasionally vary by CBSA and are not discussed in any further detail in this paper. However, Factors 3, 4 and 5 are included in model results to demonstrate that there is marginal utility in considering additional latent dimensions of urban spatial structure. But these additional latent measures must be interpreted on a market-by-market basis; for example, Factors 3 and 4 often describe suburban space as defined by transportation networks, the structure of which can vary from market-to-market. Factor 5 at the national scale identified CBGs with concentrations of poverty, but this was not consistent across all CBSAs. Thus for ease of interpretation, only Factor 1 (commercial activity

hub) and Factor 2 (urban frings) will be discussed as their subjective definitions are transferable across all markets.

Finally, although it has no effect on the conclusions of this research, the CBSA-scale PCA Factors are used in all subsequent tests of the utility of these latent urban spatial structure metrics in real estate market models. Comparison with national-scale PCA Factors reveals a marginal increase in utility for CBSA market-specific definitions of Factor 1 and Factor 2. Tests with national-scale PCA Factor 1 and 2 definitions produce similar results due to consistency of the contributing SLD variables to Factor 1 and Factor 2 across all CBSAs.

For all observations in the property transaction database the location of the property transacted results in a "factor score" for each latent measure based on the CBG in which the property sits. Descriptive statistics for the these extracted CBSA-scale PCA latent factors are shown in Table 1. Means and standard deviations are provided for each property type. Just means are provided for each CBSA. These factor scores are difficult to interpret in isolation, but they behave like an index: the mean of Factor 1 describes the mean index value of a continuous variable representing commercial hubs (relative to other CBGs within the CBSA). Factor 2 describes the mean index value of a continuous variable representing the urban fringe.

Insert Table 1 about here

A few important observations emerge in the descriptive tables. First, although the data was normalized to mean zero and standard deviation of 1, the means for Factor 1, the commercial activity factor, are well above 0 for each property product, indicating that real estate investment tends to occur in CBGs with above-average commercial activity. Unsurprisingly, office properties determined to be in a central business district exhibit the strongest signal for Factor 1. Second, all are indistinguishable from zero statistically speaking; the standard deviation for each property type exceeds the mean.

To better describe these factors, 2 shows a spatial representation of how Factor 1 and Factor 2 are continuous measures of urban spatial structure interpreted as commercial activity centers and urban fringe respectively. Atlanta and Seattle are chosen as example urban areas featuring polycentric form and geographic challenges to the concept of linear bid-rent models. Darker areas in the Factor 1 maps (a and c) represent a larger degree of commercial activity in each CBG, demonstrating that Atlanta is more polycentric than Seattle. Darker areas in the Factor 2 maps (b and d) represent CBGs on the urban fringe, with lighter areas surrounding the commercial activity centers. In combination CBGs can share characteristics that correlate with Factor 1 (activity center) and Factor 2 (fringe), resulting in a non-linear functional form of location sensitivity to real estate pricing (a non-linear bid-rent function) or risk, as will be examined here.

PCA Factors in Pricing and Risk Models

The first test of the efficacy of PCA factors as exogenous location controls is to use them as variables of interest to describe spatial variance in office and retail asset pricing and risk. Unlike housing transaction data, office and retail transactions are not as frequent, so sampling frames must expand spatial and temporal boundaries to generate sufficient observations for statistical efficiency. This expanded sampling frame adds additional variance that complicates hypotheses testing and introduces the potential for type 1 statistical error as a result of endogeneity with added spatial (or temporal) variance. Of interest here is whether the extracted latent urban form metrics (PCA factors) can measure the relationship between urban spatial structure and prices or risk.

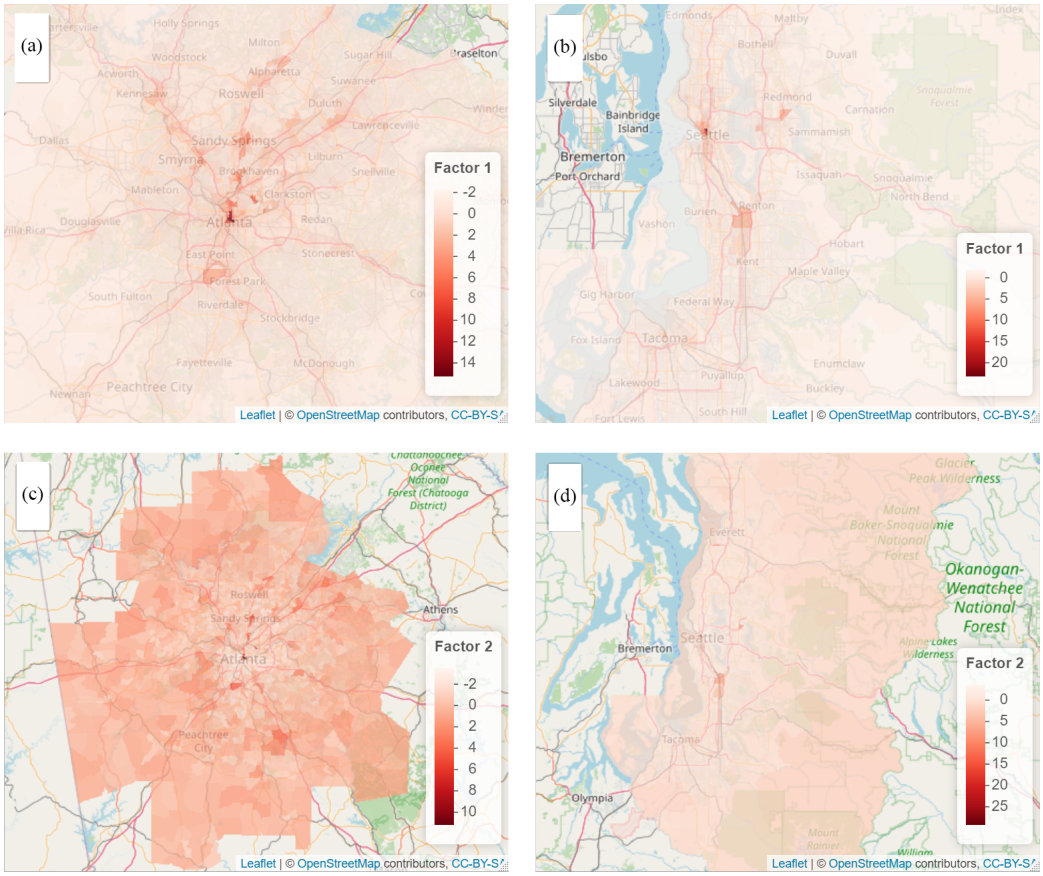

Fig. 2: Spatial representation of Factor 1 and Factor 2 "factor scores" for each CBG in Atlanta (a and c) and Seattle (b and d).

Notably, measurement of cap rates (net income yield at time of purchase) is not always possible, as it requires both a transaction price and reliable (and consistent) estimate of net income at the time of sale. Since the latter is less available than the former, there is more research on commercial real estate pricing relative to research on risk, as measured by cap rates. The ability to control for the spatial variance introduced when increasing cap rate observations by adding additional urban markets (CBSAs) would be a valuable tool in the evaluation of commercial real estate risk.

0.3 Modeling Specification

To examine the relationship between the PCA Factors and cap rates or prices, the generic specification of a model estimated using generalised method of moments (GMM) is:

$$CapRate_{ijt} = \beta_0 + \beta_1 \overrightarrow{Prop}_i + \beta_n \overrightarrow{F}_{ni} + \epsilon_{ijt} \tag{2}$$

The dependent variables will be either the transaction net income yield (cap rate), expressed as a percentage, or natural log of price per square foot as the dependent variable.[2] As only about 25% of observations include cap rates, all regressions are also run on the natural log of per square foot sales price to evaluate the sensitivity of the PCA factors to measure location value when the sample size is increased four-fold while maintaining the spatial (and temporal) variance of the sampling frame.

Models are congruent in specification with traditional commercial real estate hedonic analyses (Seiler and Walden, 2014; Gabe et al., 2021) where $\overrightarrow{Prop_i}$ is a vector of observed asset level characteristics including size and age. $\overrightarrow{N}_{ii}$ represents neighborhood characteristics not include in the SLD data such as crime and education levels. Novel to this research is that $\overrightarrow{F}_{ni}$ represents the nth factor from the principal components generated at the CBSA scale from the SLD database (described above).

These models are specified as observations of each building i, with random effects for each year t, and market j. As specified, the models facilitate addressing the research questions relative to the two problems motivating the paper: (1) ease of measuring spatial characteristics of non-linear bid-rent curves and 2) the potential for location value and price effects to be attributed to and confounded by exogenous factors that are not spatially independent. Connecting to the motivation and problems, creating Factors 1 and 2 addresses the first problem while their integration into the GMM models allows testing of both problems one and two.

Regression Results

Table 2 shows a consistent increase in pricing and decrease in cap rates across all models for Factor 1. Note that the scale of the cap rate regressions are expressed as 0 to 100. As evidence of sample validity, the per square foot premiums and cap rate premiums have comparable results. Model 1 is the reference or base model for comparative purposes.

Insert Table 2 about here

Model 2 in the sale price (LNPSF) model shows a sales premium of approximately 4.1%. The comparable reduction in cap rate from Model 2 of the cap rate models is a nominal decrease of 0.37%. The sample mean cap rate of 6.68% would then be reduced to 6.31% resulting in a 5.86% increase in price of the average property. While not identical, the comparable range is suggestive of construct validity and the reliability of signals from the smaller cap rate sample.

Factor 1 and Factor 2, the two consistent latent spatial structure variables measuring commercial activity and urban fringe respectively, reveal expected relationships between cap rates (or prices) and location. The greater the intensity of commercial activity (larger Factor 1 score), the lower the cap rate (less risk). As would be expected of a non-linear bid-rent model, the orthogonal Factor 2 exerts the opposite effect. Economically, the attraction to commercial activity centres mean the sensitivity of cap rates (or prices) is greater for Factor 1, congruent with early conceptual forms of linear bid-rent curves attracted to a central business district (Alonso, 1960). Factor 2 exhibits a weak per square foot effect and no statistical significance in the cap rate model. Given potentially disparate effects of the property types for this factor, the result here, in a model where retail and office product is combined into one model, is expected.

Specific property type models break out the cap rate based findings for the Office and Retail sectors separately. Models 1-5 in Table 3 show results for base runs including CBD controls all

[2] Additional models using the difference between observed cap rates and the RCA Cap Rate Index for the specific market were also estimated. Results converge.

but Model 2. Models 6-8 and 9-11 show sub-sample results for the Office CBD and Suburban only samples.

Insert Table 3 about here

In the base model (Model 1), control variables have expected levels of significance and effect, a nominal 1.2% cap rate reduction, on average, for being located with the RCA boundaries of a CBD. Model 2 estimates Factor 1 without an additional CBD control. Since Factor 1 and Factor 2 are normalized N(0,1) within each CBSA, the results indicate a 0.068% reduction for each standard deviation from the mean CBSA factor score.

The CBSA mean of zero was estimated based on all Census Block Groups in the entire CBSA. The mean CBD located office building Factor 1 score is 7.39 with a 7.60 standard deviation. This suggests that the average CBD office is already over seven standard deviations from the CBSA mean factor score and further exhibits rightward skew (CBD locations will generally contain greater commercial activity). This implies the impact on the mean office building would be 7.39*.041 or a 0.50 cap rate reduction. A one standard deviation shift in the CBD sample, or another 7.60 standard deviations from the mean of zero would exceed a full point, approaching the CBD dummy variable estimation.

As one of the goals of the paper is to examine the marginal impact of Factors 1 and 2 beyond traditional techniques, Models 3-5 include both Factor 1 and the CBD control. Model 3 shows a 0.033 cap rate reduction for Factor 1 along with a -0.973 reduction for traditional CBD location. Note that the mean estimate for CBD independently from Model 1 is virtually identical. The combined reduction in cap rate for the mean CBD located property would be 1.216 (0.033 Factor 1 * 7.39 mean + 0.973).

This suggests that Factor 1 does allow for increased pricing refinement beyond the simple binary mean of the CBD dummy variable. The implication here is that by measuring the extent to which any location expresses Factor 1 along the intensity spectrum (e.g., high to low economic activity), the additional benefit of Factor 1 is that it can identify pricing nuance outside traditional CBD boundaries.

Factor 2 exhibits a negative price impact (increasing cap rate). Since Factor 2 generally references suburban and or residential factors, this negative price influence for office appears in line with expectations.

Within the CBD only sample, Factor 1 shows a 0.016% decrease in cap rate per standard deviation in the factor score, or about 0.118% reduction at the mean. Authors note that the small sample size of 466 may reduce the practical applicability of this parameter estimate.

Unsurprisingly, the effect of Factor 1 is more pronounced in suburban buildings. Part of the expected utility of this Factor would be to help describe and control for polycentricity in cities with multiple commercial hubs or geographic constraints.

Easier interpretation of the the general effects of spatial location on office cap rates can be seen in 3 (a and b), which maps the combined effect of Factor 1 and Factor 2 on office cap rates in the sample markets of Atlanta and Seattle. These two maps are based on Model 4 in 3, forecasting the average spatial cap rate effect for each Census Block Group. Of note is the visual relationship between office cap rates and major transportation corridors; accessibility reduces the risk of office investments.

Table 4 shows results for the retail only sample. Similar to the Office results, Models 1-5 include all retail property types. Models 6-9 show results for strip retail only.

Insert Table 4 about here

Model 2 shows an average reduction of 0.054 for retail cap rates on Factor 1. Since retail is often located near areas of residential density, Factor 2 would be expected to positively impact

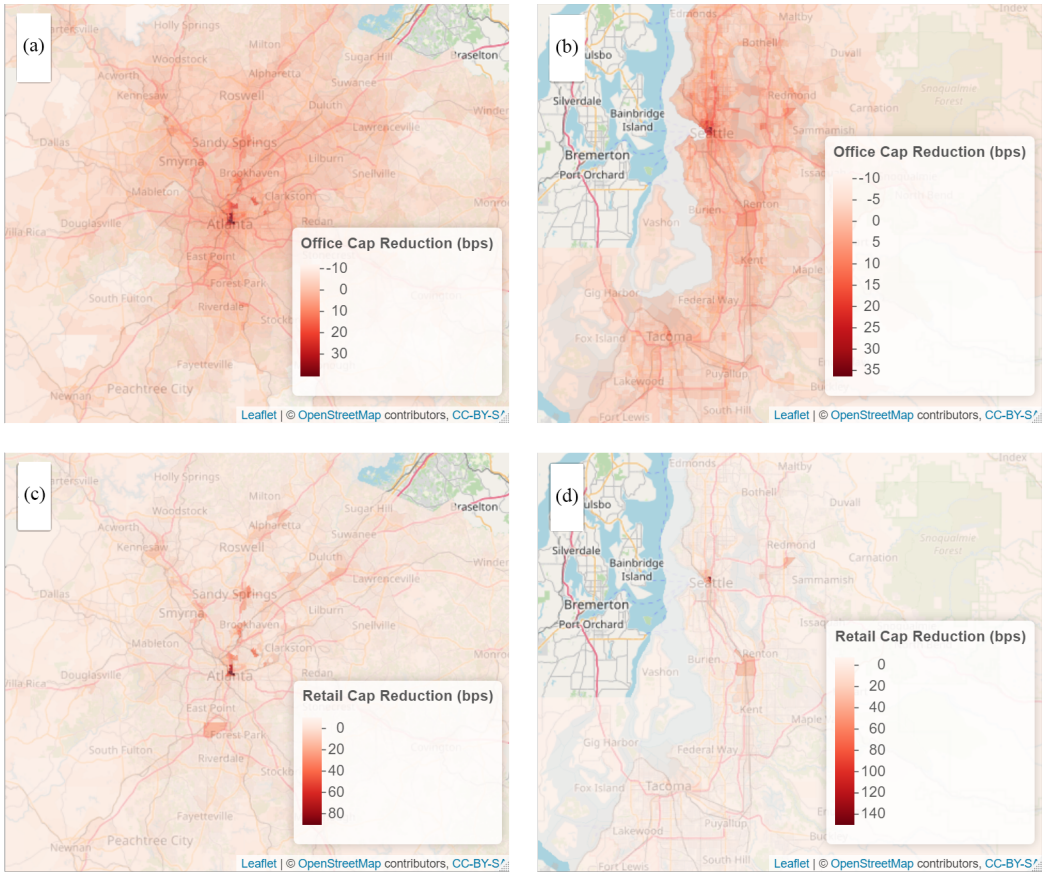

Fig. 3: Spatial variance of cap rate reductions in office markets (a and b) and retail markets (c and d) across Census Block Groups in Atlanta and Seattle. Darker areas indicate greater cap rate *reductions*, i.e. less risk. Projection based on Model 4 in 3 and 4, the specification that includes SLD Factors 1 and 2.

pricing. It does, with a 0.032 reduction in cap rate. When run concurrently with Factor 1 (Model 4), the effect of Factor 1 dominates and Factor 2 becomes statistically insignificant.

Modeling only strip retails reveals an increased importance of Factor 1 relative to mall retail. Surprisingly, Factor 2 does not exhibit statistical significance.

3 (c and d) maps the spatial variance of retail cap rates in the sample markets of Atlanta and Seattle using Model 4 in 4. Retail cap rates are relatively more sensitive to location than office, with a much larger spread in each market. There are few CBGs where retail cap rates vary, suggesting the agglomeration effects of retail attract customers (and capital) to concentrations of retail property.

Together, the office and retail modelling results contribute to the debate about simplified measurement of spatial characteristics of non-linear bid-rent curves and the potential for location value and price effects to be attributed to and confounded by exogenous factors. The creation of Factors 1 and 2 provides a relatively easy and quick pathway to capturing a range of spatial economic relationships. Their significance in the regression models suggest utility independent from

more traditional approaches where space and economic forces can be confounded when blended. Here, variation in prices and cap rates is consistent across Factor 1 and 2 and across the two asset classes. This suggests that the data reduction method and detailed micro-economic spatial data help reduce spatial bias in small sample sizes, address missing variable bias endogeneity concerns, and capture spatial economic relationships comparably to other methods.

But how does this method compare with traditional methods of spatial effects control? Out of sample testing on models of the housing market, where sample sizes are much larger to enable micro-location control, helps to describe their econometric utility of the PCA Factors further.

Out of Sample Testing: Single Family Market

The single family (SF) housing market is much larger in terms of transactions than commercial asset classes. Also, SF land uses often make up the vast majority of major metropolitan areas, especially those in the southern and western United States. As the most voluminous and spatially expansive of the real estate asset classes, the paper tests the PCA generated Factors 1 and 2 to discern both their impacts on SF home prices as well as their efficacy in improving pricing model predictive performance. This out of sample testing is useful as it provides signals against which the regression results above can be triangulated–both for construct validity and convergence.

Data for this analysis comes from the King County, Washington Tax Assessor[3]. King County is the home to Seattle and Bellevue and their immediate suburbs and is the heart of the larger Seattle-Tacoma-Everett-Bellevue Metropolitan Area. These data include all single family home sales – detached and townhomes – in the county over the January 2017 through December 2019 period, over 76,000 observations in total. Filters are applied to remove outlying observations.

Two different classes of models are used to test the impact of the SLD factors on values in the residential market – standard ordinary least squares (OLS) and random forest. Each model type uses the same set of control features, with the two SLD Factors being the variables of interest. Control variables are home size (in sq.ft), year built, home quality, home condition, bedroom count, bathroom count, lot size (logged), waterfront (binary) and combined view score (0 - 16). Time is controlled for by monthly dummy variables. The log of the home sale price is regressed against these variables along with SLD Factor 1 and Factor 2. Again, like the commercial models, the additional Factors 3-5 are included for illustrative purposes.

The linear model produces easily interpretable coefficient values, shown below in Table 5. The random forest model (RFM) does not produce such easily interpreted estimates of marginal contributions. Using a form of model agnostic interpretability – a partial dependence plot – the RFM is able to visualize the impact that the SLD variables have on predicted home values (Figure 4) and compare these against the linear coefficients from Table 5.

These two models suggest that Factors 1 and 2 have positive impacts on home prices, slighly different than was observed in office and retail markets where Factor 1 (commercial activity center) was positive and Factor 2 (urban fringe) negative. Additionally, across the model classes – OLS and random forest – the directionality of the impacts are the same, though the magnitude (the shape of the curves in Figure 4 does differ. Particularly, there is a smaller impact of Factor 2 when used in a more flexible random forest model.

Predictive Ability

Next, models examine the ability of the SLD features to properly control for spatial variation in the data. We evaluate this ability by looking into the model's predictive performance – its

[3] Data available at: www.github.com/andykrause/kingCoData

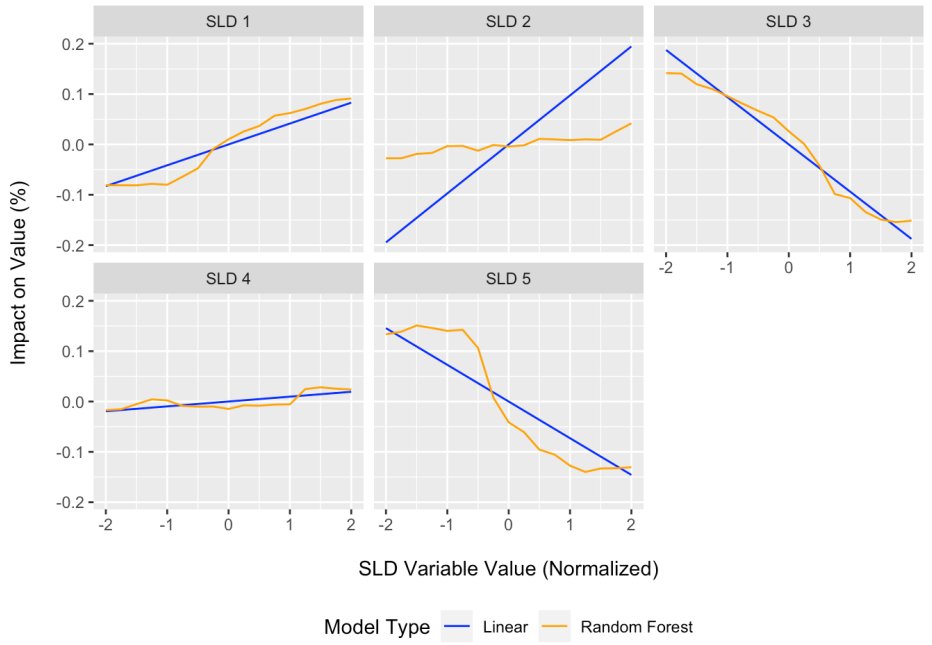

Fig. 4: Impact of SLD variables on home prices

ability to predict home prices for properties that are not in the dataset itself. To generate an out-of-sample test, an out of time approach is employed using observed transactions from the first 35 months of the data to predict the values in the final month (December 2019).

Seven different model specifications are used in order to identify the marginal impact of the SLD variables as spatial controls against other commonly used approaches to represent spatial features in house pricing models. These are:

1. Baseline: No Spatial Variables
2. Submarket: Use of fixed effects submarket binary variables
3. XY: Use of latitude, longitude and related transforms
4. SLD: Use of SLD variables
5. SLD + Submarket
6. SLD + XY
7. All: All the above

Table 6 shows results. In the linear models (left hand column), the SLD features do offer improved accuracy (Median Absolute Percentage Error, MdAPE) from the baseline model that includes no spatial variables. However, the two other approaches – submarket fixed effects and lat/long transforms – greatly outperform the SLD features. Adding the SLD features to these standard approaches results in very little improvement. This result is very similar to the utility of mobile phone tracking data as a novel location control in Bourassa et al. (2020).

For the random forest models, there are different results. The SLD feature offer a considerable improvement over the non-spatially controlled baseline model and also provide a 10% relative accuracy improvement over submarket fixed effects. Due to the flexibility of random forest models treatment of features like latitude and longitude, these standard X,Y spatial variables do outperform both the SLD and the submarket approaches. Also of note is that combining SLD

and submarket variables produces an accuracy level very similar to that of the X,Y model, suggest that the flexibility of the random forest model does allow for some complementary effects between the two. Results here help frame the problems motivating the paper and the potential for the techniques to address them.

Limitations

While the paper demonstrates Factors 1 and 2 have utility to describe urban spatial structure, these factors do not always add new statistical or economic information. For example in the single-family fixed effects specification, model fit does not always improve because market dummy variables explain much of the variance (and themselves stand in as a partial proxy for urban form).

The SLD Factors provide valuable information that isolates and extracts locational value from other exogenous variables influenced by location. In large-sample size contexts, traditional methods, such as spatial fixed effects, are comparable, and perhaps superior, particularly in linear modelling specifications. One potential area for future research would be to broaden the factor estimation from CBG to tract or some distance weighted measure of nearby CBGs. A mall is likely to be its own CBG and thus not capture the impact of nearby residential in the current modeling strategy.

The results and their limitations are consistent with both Glaeser et al. (2018b) and Bourassa et al. (2020). Big data provides opportunities for innovations in real estate and related financial economic analyses. It offers new pathways for theory to evolve, hypotheses to be tested, and signal to be split from what was once noise. In some instances, this allows for the questioning of received wisdom about human defined spatial boundaries–questions that spill forward into algorithmic fairness and other dimensions of data-defined solution alternatives where *a priori* defined models have dominated. Though the results here do not speak to this issue directly, they suggest questions that investors and researchers might want to consider in the co-production of future real estate knowledge.

Conclusions

This paper was motivated by two problems common to real estate analyses that include urban spatial structure–easy measurement of spatial characteristics of non-linear bid-rent curves and the potential for location value and price effects to be attributed to and confounded by exogenous factors–especially treatments without random spatial distributions. Motivated by these problems and the ever expanding universe of micro-economic data, the paper explores the use of Principal Component Analysis (PCA) to extract two latent factors of real estate location from 90 measures of urban form by the Environmental Protection Agency at the Census Block Group scale.

These latent variables describe, in continuous functional form, the urban and exurban intensity of each CBG in each CBSA. These two factors represent the utility functions underlying bid-rent curves for location value within an urban system. They provide a simplified and alternative methodological pathway to other econometric and geospatial advances such as fixed effects and autocorrelation techniques. Factors 1 and 2 also create the potential to evaluate, at a smaller scale and on a more detailed basis, the relationships between urban spatial attributes and prices/capitalization rates than current practice, which largely consists subjective definitions of *submarkets* or definitions of Central Business Districts. In this context, Factors 1 and 2 were incorporated into hedonic regression models analyzing a sample of transaction prices and capitalization rates from Real Capital Analytics detailing office and retail assets in more than 35 U.S. Core Based Statistical Areas. The results from these models and out of sample tests

indicates two contributions literature and practice. The first contribution is a proof of concept. Models indicate that across each of the asset classes - housing, office, and retail - both transaction prices and capitalization rates are both statistically and economically sensitive to variation in the continuous form specifications of each PCA factor. This leads to market structure insights, notably that retail risk is more location-sensitive than office, which has less spatial heterogeneity. Accuracy of predictions using PCA Factors in housing price models show similarities with current techniques of spatial control.

Practically, this first contribution means that real estate professionals can evaluate spatial bias in capital allocation decisions. With an exogenous and non-linear measure of urban spatial structure, savvy investors can investigate deviations from the expected spatial risk premium. For example cap rates in a CBG higher than the spatial model would predict suggest other exogenous or market effects that are not location-based, such as socioeconomic factors or industry agglomerations, are increasing perceived risk. Or such deviation could reflect a pricing inefficiency caused by imperfect information on urban form productivity.

Second, the use of latent PCA factors of urban spatial structure "big data" can control for spatial bias in real estate models similar to current methods but the exogenous nature of these latent factors fills a niche to control for spatial bias in models with small sample sizes. While out-of-sample forecasting accuracy suggests these factors are, at best, comparable to traditional large-sample size methods of location control (i.e. submarket fixed effects), the office cap rate models demonstrated that these factors improve on existing location controls for small-sample sizes, such as distance-to-CBD measures.

Practically, this second contribution opens up the potential to disentangle spatial endogeneity in contexts, like cap rate models, where large sample sizes are not possible, or when treatments of interest are not distributed spatially at random. For example, health and wellness building certifications, like the WELL building standard, are only featured on a small fraction of assets traded in a market in any given period; to evaluate the effect of such treatment on prices, the sampling frame must expand spatially (and temporally), introducing variance that may correlate with the treatment and thus produce Type 1 error through spaital endogeneity. An exogenous measure to control for the introduced spatial variance could theoretically reduce probability of a false positive signal for the subject treatment.

Naturally, as a step forward, the results raise a useful set of questions that future research might examine–questions about a priori defined rules within spatial analysis, their application to the diverse landscape of commercial real estate, and the potential (and attendant issues) of algorithmic or missing variable bias.

Acknowledgements: This paper was made possible with support from the Real Estate Research Institute and Real Capital Analytics. A sincere thank you to Robert White, Jim Costello, and Matthew Benz for data and insights about the shape and functionality of urban property markets. Similarly, a significant thank you to Dr. Mark Eppli and Dr. Elaine Worzala for their guidance. As mentors, their stewardship of the project was invaluable. We also thank the Board and staff at RERI for their support, both intellectual and financial.

Table 1: Descriptive. Standard deviation for property types are shown below the mean. Only means are shown for CBSA for brevity.

		Factor								
Variable	N	1	2	3	4	5	Age	Sale PSF	Sale Price (000)	Square Ft (000)
All Property	40,871	1.424	0.801	-0.209	-0.116	-0.134	35	266	15,739	85
		2.870	2.136	2.656	1.652	3.314	27	381	55,318	157
Flex	4,247	1.370	0.798	-0.255	-0.173	-0.124	32	142	8,946	78
		1.934	1.996	2.160	1.749	3.627	19	135	17,689	117
Mall & Other	6,810	1.338	0.595	0.243	-0.076	-0.033	45	575	12,088	35
		2.707	1.854	2.371	1.339	2.582	38	728	51,195	109
Office - CBD	1,920	7.386	3.655	-4.021	0.177	-1.240	65	484	85,051	190
		7.601	5.529	7.614	2.700	6.840	37	454	198,105	296
Office - Sub	10,428	1.361	0.719	-0.162	-0.064	0.034	31	232	17,718	86
		1.996	1.701	2.057	1.508	2.911	20	194	39,323	145
Strip	7,271	0.487	0.546	0.402	0.064	0.038	26	242	10,973	60
		1.202	1.304	1.121	1.346	1.690	19	190	17,445	77
Warehouse	10,195	1.115	0.668	-0.259	-0.354	-0.290	37	122	9,330	117
		1.780	1.665	1.818	1.839	3.777	22	136	14,377	194
Atlanta	1,334	0.828	0.990	0.245	-0.190	0.831	26	153	13,700	119
Austin	492	1.300	0.947	0.388	0.088	-0.490	25	264	17,338	77
Baltimore	527	0.621	2.528	-1.054	-0.336	-2.025	34	158	12,522	105
Boston Metro	1,258	0.745	1.584	0.107	-0.307	0.729	46	244	23,704	98
Charlotte	557	0.962	1.020	-0.125	-0.213	0.923	24	181	13,832	107
Chicago	2,259	1.585	1.435	-0.283	-0.371	0.818	35	183	15,876	119
Cincinnati	239	0.971	1.799	-0.230	0.020	-1.923	27	118	10,867	130
Columbus	338	0.842	1.482	0.017	0.121	-0.328	25	114	11,101	168
DC Metro	1,354	1.486	1.536	-0.595	-0.028	-0.793	36	332	27,537	97
Dallas	1,934	1.601	0.987	-0.310	-0.398	1.817	26	186	13,840	106
Denver	1,032	1.361	0.990	-0.130	-0.446	-0.243	30	205	14,198	79
Detroit	589	1.582	0.922	0.128	-0.693	1.656	30	129	8,855	107
Honolulu	121	1.194	-0.827	0.571	0.003	0.416	39	479	17,421	51
Houston	1,289	1.042	0.790	0.130	0.171	-0.088	24	178	13,928	97
Indianapolis	425	0.923	1.454	-0.028	-0.009	0.535	27	129	9,433	123
Jacksonville	238	1.104	1.556	0.599	-0.104	-1.352	24	173	12,229	138
Kansas City	360	1.495	1.350	-1.143	-0.253	-2.232	32	162	13,613	119
LA Metro	6,352	1.801	-0.636	-0.647	-0.105	-1.321	38	302	12,656	59
Las Vegas	639	1.550	-0.595	1.091	-0.184	0.723	18	242	16,134	69
Memphis	270	0.908	1.696	0.145	-0.520	3.009	26	119	8,868	190
Miami/So Fla	1,745	1.597	-0.827	-0.692	0.438	-1.502	33	287	13,343	66
Minneapolis	733	0.897	1.690	-0.067	-0.372	-1.511	33	153	11,627	97
NYC Metro	4,279	1.453	1.060	-0.286	-0.201	0.592	61	572	25,719	68
Nashville	456	1.603	1.522	0.508	-0.106	0.624	31	201	11,575	101
Norfolk	203	2.193	1.098	-0.547	0.212	-2.290	26	147	9,575	91
Orlando	578	1.138	1.274	-0.271	-0.054	-0.582	25	189	11,164	89
Philly Metro	1,058	-0.447	2.233	-0.176	-0.346	1.575	37	172	13,732	105
Phoenix	1,415	1.404	1.035	-0.291	0.102	-0.697	22	183	11,972	78
Pittsburgh	170	1.229	1.794	-0.195	0.011	-0.582	31	174	17,546	145
Portland	540	1.683	1.122	0.105	-0.202	1.329	40	194	13,176	78
Raleigh/Durham	385	1.191	1.133	0.072	0.205	0.257	24	206	12,848	84
Richmond	257	1.584	1.290	-0.075	0.183	-1.361	27	122	8,324	91
SF Metro	2,601	1.820	0.640	-0.697	0.268	-0.803	45	355	21,114	68
Sacramento	664	1.525	1.100	-0.357	0.158	1.214	28	175	8,403	67
Salt Lake City	406	1.787	1.329	-1.169	0.705	-1.669	28	142	10,837	78
San Antonio	427	0.975	1.027	0.155	0.252	-0.540	23	188	9,960	72
San Diego	1,026	2.309	1.618	1.954	-0.750	1.694	31	272	11,790	49
Seattle	1,158	1.452	0.704	0.766	-0.255	-0.901	36	285	20,324	71
St Louis	601	1.554	1.567	-0.245	0.034	-2.391	31	119	8,425	104
Tampa	562	1.609	1.116	-0.368	-0.156	0.944	28	181	12,805	87

Table 2: All Property Type Hedonic Regression Estimates:
Each model shows results from a GMM mixed effect regression with a dependent variable of the natural log of average rent PSF or a capitalization rate (*100) respectively. Factors are by CBSA and normalized with a mean zero and standard deviation of one. Warehouse is the omitted property type category. All models include mixed effect controls for CBSA and year. ***, ** and * indicate significance at 99%. 95% and 90% levels, respectively. Standard errors shown beneath estimate are clustered at market level.

Dependent	LNPSF				Cap Rate * 100			
Variable	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Intercept	7.951	8.046	8.048	8.029	5.401	5.283	5.281	5.285
	181.081	185.679	185.777	185.685	27.792	27.176	27.16	27.17
lnsf	-0.254***	-0.266***	-0.265***	-0.264***	0.024*	0.036***	0.036***	0.037***
	-88.838	-93.79	-93.591	-93.458	1.794	2.725	2.692	2.731
lnage	-0.385***	-0.367***	-0.366***	-0.360***	0.492***	0.480***	0.479***	0.475***
	-29.15	-28.158	-28.09	-27.739	9.814	9.589	9.581	9.479
lnage2	0.048***	0.041***	0.041***	0.039***	-0.081***	-0.075***	-0.075***	-0.074***
	20.093	17.543	17.355	16.495	-8.039	-7.523	-7.495	-7.311
Factor 1		0.041***	0.044***	0.058***		-0.037***	-0.040***	-0.046***
		34.374	33.176	34.407		-7.189	-6.84	-6.227
Factor 2			-0.009***	-0.007***			0.008	0.009
			-5.026	-3.785			1.024	1.125
Factor 3				0.021***				-0.006
				12.383				-0.747
Factor 4				-0.001				0.007
				-0.644				0.813
Factor 5				0.002*				-0.007
				1.755				-1.377
Strip	0.634***	0.647***	0.647***	0.642***	0.876***	0.875***	0.876***	0.872***
	65.648	68.004	67.994	67.438	17.964	17.996	18.004	17.875
Flex	0.150***	0.134***	0.134***	0.130***	0.784***	0.806***	0.806***	0.807***
	13.514	12.222	12.22	11.907	11.974	12.323	12.321	12.33
Mall & Other	0.964***	0.934***	0.933***	0.921***	-0.059	-0.029	-0.027	-0.026
	91.747	89.929	89.816	88.403	-1.05	-0.51	-0.48	-0.468
Office - CBD	1.229***	0.971***	0.977***	0.968***	-0.451***	-0.138	-0.147*	-0.152*
	79.881	57.406	57.637	56.901	-6.253	-1.642	-1.735	-1.773
Office - Sub	0.608***	0.589***	0.588***	0.581***	0.894***	0.925***	0.927***	0.927***
	70.955	69.613	69.455	68.564	19.16	19.804	19.829	19.759
AIC	67,245	66,084	66,070	65,920	26,873	27,997	28,004	28,024
SIC	67,241	66,080	66,066	65,916	26,871	27,993	28,000	28,020
Model N	38,198	38,195	38,195	38,195	8,192	8,190	8,190	8,190

Table 3: Office Hedonic Regression Estimates by Cap Rate:
Each model shows results from a GMM mixed effect regression with a dependent variable of the capitalization rate (*100). Models 1-5 include all office observations, Models 6-8 include only those in a CBD and Models 9-11 only suburban. Factors are by CBSA and normalized with a mean zero and standard deviation of one. All models include mixed effect controls for CBSA and year. ***, ** and * indicate significance at 99%. 95% and 90% levels, respectively. Standard errors shown beneath estimate are clustered at market level.

						CBD Only			Suburban Only		
Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Intercept	7.042	7.038	6.802	6.797	6.799	8.741	8.721	8.565	6.224	6.215	6.173
	19.892	19.441	19.117	19.11	19.101	9.213	9.097	8.637	15.952	15.916	15.77
lnsf	-0.061**	-0.071***	-0.035	-0.035	-0.035	-0.178***	-0.178***	-0.162***	0.01	0.01	0.012
	-2.552	-2.871	-1.42	-1.43	-1.407	-3.427	-3.413	-3.044	0.375	0.354	0.427
lnage	0.772***	1.041***	0.772***	0.771***	0.763***	0.286	0.287	0.304	0.868***	0.869***	0.863***
	7.011	9.592	7.041	7.038	6.954	1.346	1.347	1.424	6.755	6.762	6.715
lnage2	-0.126***	-0.184***	-0.124***	-0.124***	-0.121***	-0.049	-0.049	-0.05	-0.142***	-0.142***	-0.139***
	-5.973	-8.896	-5.887	-5.87	-5.741	-1.337	-1.335	-1.358	-5.461	-5.463	-5.358
Factor 1		-0.068***	-0.033***	-0.041***	-0.050***	-0.016**	-0.015	-0.03	-0.082***	-0.083***	-0.086***
		-11.608	-4.879	-5.139	-4.513	-2.299	-1.247	-1.349	-5.208	-5.221	-4.943
Factor 2				0.020*	0.020*		-0.003	-0.006		0.011	0.018
				1.907	1.759		-0.157	-0.327		0.563	0.827
Factor 3					-0.015			-0.006			0.001
					-1.252			-0.248			0.086
Factor 4					0.016			-0.032			0.053***
					1.055			-1.336			2.634
Factor 5					-0.001			-0.005			0.011
					-0.154			-0.496			0.935
CBD	-1.216***		-0.973***	-0.992***	-1.008***						
	-14.723		-10.121	-10.27	-10.081						
AIC	10,341	10,319	10,423	10,326	10,329	2,951	2,957	2,973	8,515	8,520	8,533
SIC	10,335	10,313	10,417	10,320	10,323	2,945	2,951	2,967	8,511	8,516	8,529
Model N	2,621	2,621	2,621	2,621	2,621	466	466	466	2,155	2,155	2,155

Table 4: Retail Hedonic Regression Estimates by Cap Rate:
Each model shows results from a GMM mixed effect regression with a dependent variable of the capitalization rate (*100). Models 1-5 include all retail observations, Models 6-10 include only those defined as strip mall retail. Factors are by CBSA and normalized with a mean zero and standard deviation of one. All models include mixed effect controls for CBSA and year. ***, ** and * indicate significance at 99%. 95% and 90% levels, respectively. Standard errors shown beneath estimate are clustered at market level.

						Strip Retail Only				
Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Intercept	4.403	4.409	4.392	4.406	4.405	6.032	6.042	6.035	6.046	6.058
	20.168	20.255	20.124	20.237	20.204	20.784	20.869	20.756	20.852	20.873
lnsf	0.161***	0.165***	0.165***	0.167***	0.167***	0.054**	0.057**	0.055**	0.057**	0.059**
	8.959	9.223	9.145	9.251	9.271	2.261	2.375	2.278	2.386	2.438
lnage	0.450***	0.421***	0.447***	0.421***	0.419***	0.585***	0.577***	0.585***	0.577***	0.566***
	7.855	7.327	7.793	7.325	7.285	6.942	6.858	6.935	6.854	6.707
lnage2	-0.086***	-0.076***	-0.086***	-0.077***	-0.076***	-0.088***	-0.084***	-0.088***	-0.085***	-0.081***
	-7.312	-6.408	-7.277	-6.428	-6.346	-4.93	-4.742	-4.937	-4.749	-4.533
Factor 1		-0.054***		-0.051***	-0.056***		-0.077***		-0.077***	-0.085***
		-5.023		-4.504	-4.311		-3.655		-3.654	-3.81
Factor 2			-0.032**	-0.011	-0.01			-0.006	-0.006	-0.01
			-2.337	-0.737	-0.681			-0.289	-0.291	-0.453
Factor 3					-0.003					-0.034
					-0.242					-1.348
Factor 4					0.003					0.02
					0.203					1.184
Factor 5					-0.01					0.008
					-0.954					0.485
Strip	0.792***	0.764***	0.789***	0.764***	0.762***					
	19.245	18.446	19.19	18.456	18.363					
AIC	11,866	11,843	11,862	11,849	11,869	8,003	7,142	7,155	7,148	7,164
SIC	11,864	11,841	11,860	11,847	11,867	7,999	7,138	7,151	7,144	7,160
Model N	3,832	3,830	3,830	3,830	3,830	2,273	2,272	2,272	2,272	2,272

Table 5: Single Family Regression Estimates. This table shows results of the Factors on King County Washington single family home sales from January 2017 through December 2019.

	Estimate	Std. Error	t value	Pr(>|t|)
SLD_1	0.0415	0.0020	20.36	0.0000
SLD_2	0.0973	0.0021	47.43	0.0000
SLD_3	-0.0938	0.0016	-58.78	0.0000
SLD_4	0.0097	0.0013	7.27	0.0000
SLD_5	-0.0729	0.0013	-56.41	0.0000

Table 6: Single Family Housing Random Forest Models. This table compares results from the random forest models to the ordinary least squares models.

	spec	OLS	Random Forest
1	base	0.20	0.18
2	subm	0.13	0.10
3	xy	0.15	0.08
4	sld	0.19	0.09
5	sld+subm	0.13	0.09
6	sld+xy	0.15	0.08
7	sld + xy + subm	0.13	0.08

References

Abdi, H. and L. J. Williams (2010): "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, 2, 433–459.

Ahlfeldt, G. M. and N. Wendland (2013): "How polycentric is a monocentric city? centers, spillovers and hysteresis," *Journal of Economic Geography*, 13, 53–83.

Alonso, W. (1960): "A theory of the urban land market," *Papers in Regional Science*, 6, 149–157.

Alonso, W. (1968): "Urban and regional imbalances in economic development," *Economic development and cultural change*, 17, 1–14.

An, X. and G. Pivo (2020): "Green buildings in commercial mortgage-backed securities: the effects of leed and energy star certification on default risk and loan terms," *Real Estate Economics*, 48, 7–42.

Anas, A., R. Arnott, and K. A. Small (1998): "Urban spatial structure," *Journal of economic literature*, 36, 1426–1464.

Anas, A. and I. Kim (1996): "General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration," *Journal of Urban Economics*, 40, 232–256.

Barreca, A., R. Curto, and D. Rolando (2020): "Urban vibrancy: an emerging factor that spatially influences the real estate market," *Sustainability*, 12, 346.

Bialkowski, J., S. Titman, and G. J. Twite (2021): "The determinants of office rents and yields: The international evidence," *Working Paper*.

Bond, S. A. and A. Devine (2016): "Certification matters: is green talk cheap talk?" *The Journal of Real Estate Finance and Economics*, 52, 117–140.

Bourassa, S. C., E. Cantoni, and M. Hoesli (2007): "Spatial dependence, housing submarkets, and house price prediction," *The Journal of Real Estate Finance and Economics*, 35, 143–160.

Bourassa, S. C., M. Hoesli, L. Merlin, and J. Renne (2020): "Big data, accessibility, and urban house prices," *Urban Studies*.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–79.

Cotter, J. and R. Roll (2015): "A comparative anatomy of residential reits and private real estate markets: returns, risks and distributional characteristics," *Real Estate Economics*, 43, 209–240.

Eberts, R. W. and D. P. McMillen (1999): "Agglomeration economies and urban public infrastructure," *Handbook of regional and urban economics*, 3, 1455–1495.

Ewing, R. and R. Cervero (2010): "Travel and the built environment: a meta-analysis," *Journal of the American planning association*, 76, 265–294.

Fisher, G., E. Steiner, F. F. Ventures, S. Titman, and A. Viswanathan (2020): "How does property location influence investment risk and return?" Technical report, Working Paper.

Freeman, L. C. (1977): "A set of measures of centrality based on betweenness," *Sociometry*, 35–41.

Fuerst, F. and P. McAllister (2011): "Green noise or green value? measuring the effects of environmental certification on office values," *Real Estate Economics*, 39, 45–69.

Gabe, J., S. Robinson, and A. Sanderford (2021): "Willingness to pay for attributes of location efficiency," *Journal of Real Estate Finance and Economics*.

Glaeser, E. L., H. Kim, and M. Luca (2018a): "Nowcasting gentrification: using yelp data to quantify neighborhood change," in *AEA Papers and Proceedings*, volume 108, volume 108, 77–82.

Glaeser, E. L., S. D. Kominers, M. Luca, and N. Naik (2018b): "Big data and big cities: The promises and limitations of improved measures of urban life," *Economic Inquiry*, 56, 114–137.

Gordon-Larsen, P., M. C. Nelson, P. Page, and B. M. Popkin (2006): "Inequality in the built environment underlies key health disparities in physical activity and obesity," *Pediatrics*, 117, 417–424.

Hansen, W. G. (1959): "How accessibility shapes land use," *Journal of the American Institute of planners*, 25, 73–76.

Heikkila, E., P. Gordon, J. I. Kim, R. B. Peiser, H. W. Richardson, and D. Dale-Johnson (1989): "What happened to the cbd-distance gradient?: land values in a policentric city," *Environment and planning A*, 21, 221–232.

Heinig, S. and A. Nanda (2018): "Measuring sentiment in real estate–a comparison study," *Journal of Property Investment & Finance*.

Kok, N., P. Monkkonen, and J. M. Quigley (2014): "Land use regulations and the value of land and housing: An intra-metropolitan analysis," *Journal of Urban Economics*, 81, 136–148.

Kuang, C. (2017): "Does quality matter in local consumption amenities? an empirical investigation with yelp," *Journal of Urban Economics*, 100, 1–18.

Limehouse, F. and R. E. McCormick (2011): "Impacts of central business district location: A hedonic analysis of legal service establishments," *US Census Bureau Center for Economic Studies Working Paper No. CES*, 11–21.

Lucas, R. E. and E. Rossi-Hansberg (2002): "On the internal structure of cities," *Econometrica*, 70, 1445–1476.

McMillen, D. P. and J. F. McDonald (1997): "A nonparametric analysis of employment density in a polycentric city," *Journal of Regional Science*, 37, 591–612.

Muth, R. F. (1975): *Urban economic problems*, HarperCollins Publishers.

Naik, N., R. Raskar, and C. A. Hidalgo (2016): "Cities are physical too: Using computer vision to measure the quality and impact of urban appearance," *American Economic Review*, 106, 128–32.

Peng, L. (2016): "The risk and return of commercial real estate: A property level analysis," *Real Estate Economics*, 44, 555–583.

Pivo, G. and J. D. Fisher (2010): "Income, value and returns in socially responsible office properties," *Journal of Real Estate Research*, 32, 243–270.

Pivo, G. and J. D. Fisher (2011): "The walkability premium in commercial real estate investments," *Real Estate Economics*, 39, 185–219.

Porta, S., E. Strano, V. Iacoviello, R. Messora, V. Latora, A. Cardillo, F. Wang, and S. Scellato (2009): "Street centrality and densities of retail and services in bologna, italy," *Environment and Planning B: Planning and design*, 36, 450–465.

Ramsey, K. and A. Bell (2014): "The smart location database: A nationwide data resource characterizing the built environment and destination accessibility at the neighborhood scale," *Cityscape*, 16, 145.

Ringnér, M. (2008): "What is principal component analysis?" *Nature biotechnology*, 26, 303–304.

Roback, J. (1982): "Wages, rents, and the quality of life," *Journal of political Economy*, 90, 1257–1278.

Rosen, S. (1974): "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy*, 82, 34–55.

Sabidussi, G. (1966): "The centrality index of a graph," *Psychometrika*, 31, 581–603.

Saiz, A. (2010): "The geographic determinants of housing supply," *The Quarterly Journal of Economics*, 125, 1253–1296.

Seiler, M. and E. Walden (2014): "Lender characteristics and the neurological reasons for strategic mortgage default," *Journal of Real Estate Research*, 36, 341–362.

Sevtsuk, A. and C. Ratti (2010): "Does urban mobility have a daily routine? learning from the aggregate data of mobile networks," *Journal of Urban Technology*, 17, 41–60.

Sivitanides, P., R. Torto, and W. Wheaton (2003): "Real estate market fundamentals and asset pricing," *The Journal of Portfolio Management*, 29, 45–53.

Song, Y. and G.-J. Knaap (2004): "Measuring urban form: Is portland winning the war on sprawl?" *Journal of the American Planning Association*, 70, 210–225.

Titman, S. et al. (1985): "Urban land prices under uncertainty," *American Economic Review*, 75, 505–514.

Wheaton, B. (1979): *Monocentric models of urban land use: Contributions and criticism, in Current Issues in Urban Economics, P. Mieszkowski and M. Straszheim, Eds.*, Johns Hopkins Press.

Wheaton, W. C. (1974): "A comparative static analysis of urban spatial structure," *Journal of Economic Theory*, 9, 223–237.

Wieand, K. (1987): "An extension of the monocentric urban spatial equilibrium model to a multicenter setting: The case of the two-center city," *Journal of Urban Economics*, 21, 259–271.

Wold, S., K. Esbensen, and P. Geladi (1987): "Principal component analysis," *Chemometrics and intelligent laboratory systems*, 2, 37–52.

Appendix

Table 7: Five "D" Descriptive Statistics:

Variable	Source	Definition
Density_CBG_Residential	EPA SLD	Gross residential density (HU/acre) on unprotected land
Density_Emp_Retail_8	EPA SLD	Gross retail (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Office_8	EPA SLD	Gross office (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Ind_8	EPA SLD	Gross industrial (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Service_8	EPA SLD	Gross service (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Ent_8	EPA SLD	Gross entertainment (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Edu_8	EPA SLD	Gross education(8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Health_8	EPA SLD	Gross health care (8-tier) employment density (jobs/acre) on unprotected land
Density_Emp_Public_8	EPA SLD	Gross pu8blic (8-tier) employment density (jobs/acre) on unprotected land
Design_CBG_AutoLink	EPA SLD	Network density in terms of facility miles of auto-oriented links per square mile
Design_CBG_MultiLinks	EPA SLD	Network density in terms of facility miles of multi-modal links per square mile
Design_CBG_PedestrianLink	EPA SLD	Network density in terms of facility miles of pedestrian-oriented links per square mile
Design_CBG_IntersectionWeighted	EPA SLD	Street intersection density (weighted, auto-oriented intersections eliminated)
Design_CBG_AutoInstersection_PSM	EPA SLD	Intersection density in terms of auto-oriented intersections per square mile
Design_CBG_Multi3Leg_PSM	EPA SLD	Intersection density in terms of multi-modal intersections having three legs per square mile
Design_CBG_Multi4Leg_PSM	EPA SLD	Intersection density in terms of multi-modal intersections having four or more legs per square mile

Variable	Source	Definition
Design_CBG_Ped3Leg_PSM	EPA SLD	Intersection density in terms of pedestrian-oriented intersections having three legs per square mile
Design_CBG_Ped4Leg_PSM	EPA SLD	Intersection density in terms of pedestrian-oriented intersections having four or more legs per square mile
Design_CBG_Park	GIS	Euclidean distance to nearest park polygon edge, derived using Near tool in ArcGIS
Design_Apartment_ParkAccess	CDC	*apartment data only. CDC-based park accessibility score. Processed with Spatial Join function in ArcGIS.
Destination_Emp_Bike	ACS Census	MEANS OF TRANSPORTATION TO WORK: Bicycle: Workers 16 years and over – (Estimate)
Destination_Emp_Walk	ACS Census	MEANS OF TRANSPORTATION TO WORK: Walked: Workers 16 years and over – (Estimate)
Destination_CBG_Auto_Jobs45Min	EPA SLD	Jobs within 45 minutes auto travel time, timedecay (network travel time) weighted
Destination_CBG_RegAccess_Auto	EPA SLD	Proportional Accessibility to Regional Destinations - Auto: Working age population accessibility expressed as a ratio of total CBSA accessibility
Destination_CBG_RegCntrlty_Auto	EPA SLD	Regional Centrality Index – Auto: CBG D5ce score relative to max CBSA D5ce score
Destination_CBG_Walkability	EPA	Index from National Walkability data
Destination_Emp_45Minsplus	ACS Census	Derived from US Census Bureau's ACS data for all commute trips 45 minutes or longer in temporal duration
Destination_Emp_30_45Mins	ACS Census	Derived from US Census Bureau's ACS data for all commute trips of temporal duration between 30 and 45 minutes
Destination_Emp_10_30Mins	ACS Census	Derived from US Census Bureau's ACS data for all commute trips between 10 and 30 minutes in temporal duration
Destination_Emp_10Mins	ACS Census	Derived from US Census Bureau's ACS data for all commute trips up to 10 minutes in temporal duration
Distance_Emp_Trnst_QtrMile	EPA SLD	Proportion of CBG employment within 1/4 mile of fixed-guideway transit stop
Distance_Emp_Trnst_HalfMile	EPA SLD	Proportion of CBG employment within 1/2 mile of fixed-guideway transit stop
Distance_CBG_TrnstFreq_PSM	EPA SLD	Aggregate frequency of transit service (D4c) per square mile
Diversity_CBG_Edu_College_Some	ACS Census	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER: Some college, 1 or more years, no degree: Population 25 years and over – (Estimate)
Diversity_CBG_Edu_College_Trade	ACS Census	EDUCATIONAL ATTAINMENT FOR THE POPULATION 25 YEARS AND OVER: Professional school degree: Population 25 years and over – (Estimate)
Diversity_CBG_Edu_Bach_Assoc	ACS Census	Derived from US Census Bureau's ACS data for bachelor's and associates degrees received.
Diversity_CBG_Edu_Graduate	ACS Census	Derived from US Census Bureau's ACS data for graduate degrees received.
Diversity_CBG_Pop_Mean_Income	ACS Census	HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2015 INFLATION-ADJUSTED DOLLARS): Total: Households – (Estimate)
Diversity_CBG_Owner_Occupied	ACS Census	TENURE: Owner occupied: Occupied housing units – (Estimate)
Diversity_CBG_PercentLowWage	EPA SLD	% LowWageWk of total #workers in a CBG (home location), 2010
Diversity_Emp_JobsPerHousehold	EPA SLD	Jobs per household
Diversity_Emp_Entropy_8	EPA SLD	8-tier employment entropy (denominator set to observed employment types in the CBG)
Diversity_CBG_TripEquilibrium	EPA SLD	Trip productions and trip attractions equilibrium index; the closer to one, the more balanced the trip making

Variable	Source	Definition
Diversity_Region_Emp_Diversity	EPA SLD	Regional Diversity. Standard calculation based on population and total employment: Deviation of CBG ratio of jobs/pop from regional average ratio of jobs/pop
Diversity_Region_Emp_WkrsPerJob	EPA SLD	Household Workers per Job, as compared to the region: Deviation of CBG ratio of household workers/job from regional average ratio of household workers/job
Diversity_Emp_WorkersPerJob	EPA SLD	Household Workers per Job, by CBG
Diversity_Emp_Equilibrium	EPA SLD	Household Workers per Job Equilibrium Index; the closer to one the more balanced the resident workers and jobs in the CBG.

Notes: This table identifies the source and brief definition of the five "D," Density, Design, Destination Accessibility, Distance to Transit and Diversity, variables used in this analysis.