Applications of Computer Vision in Evaluating the Effects of New Housing Projects

by

You Xuan Thung

B.A., University of Cambridge (2021)

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computational Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author .....
Center for Computational Science and Engineering
August 5, 2022

Certified by .....
Fábio Duarte
Principal Research Scientist, Senseable City Lab
Thesis Supervisor

Certified by .....
Carlo Ratti
Professor of the Practice
Director, Senseable City Lab
Thesis Reader

Accepted by .....
Youssef M. Marzouk
Professor of Aeronautics and Astronautics
Co-Director, Center for Computational Science and Engineering

Applications of Computer Vision in Evaluating the Effects of New Housing Projects

by

You Xuan Thung

Submitted to the Center for Computational Science and Engineering
on August 5, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Computational Science and Engineering

Abstract

Cities are laden with visual clues. Tapping on the large volume of street view imagery (SVI) made available in the last decade, we investigate how modern computer vision tools can characterize the visual quality and linguistic diversity of cities and leverage on these novel metrics to study the impact of new housing projects.

Streets form a public space and how they look plays an important role in shaping how walkable they are, how safe people perceive them to be, and the general quality of living in the urban environment. To provide useful metrics to quantify the quality of streets, we construct a scalable process with state-of-the-art machine learning models to generate second-order metrics which capture both physical and perceptual features in an urban environment. Recognizing that the abundance of linguistic features littered across streetscapes gives us clues about underlying individual and social preferences in streetscapes, we also seek to quantify the linguistic diversity in cities. To that end, we construct a language detection model supporting English, Swedish, Arabic and Chinese that outperforms existing optical character recognition (OCR) tools. We evaluate visual interpretability with gradient-weighted class activation maps (Grad-CAM) and find that our model is both accurate and interpretable. We apply these tools to our case study of Stockholm and find intuitive spatiotemporal characterizations of the city. We also advance the application of these metrics by using them in a difference-in-difference (DID) setting to study the effects of newly completed housing projects on the built environment. We find that these projects generate spillover effects, as evident in the increase in enclosure and linguistic diversity in their immediate surroundings.

Thesis Supervisor: Fábio Duarte
Title: Principal Research Scientist, Senseable City Lab

Acknowledgments

It has been slightly more than a year since I sat behind a webcam at 1am (in true pandemic style) to share about my undergraduate research with members of the Senseable City Lab. Short as it may have been, the last year has been an incredible journey of growth and nothing short of memorable.

I would first like to thank Fábio, my advisor, for bringing me into Senseable and for readily agreeing to be my advisor. Thank you for putting your trust in me, only a recent graduate in Economics, to pursue my interests in machine learning and urban data analytics, for helping to reshape and refine my muddled ideas, and for providing so many avenues of growth along the way. Thank you, also to my thesis reader, Carlo for providing the incisive feedback needed to sharpen the work here.

I remember being somewhat intimidated about making my first presentation at our fortnightly research meetings, but in hindsight, it has been an incredible opportunity to bounce ideas with an extremely constructive and collaborative community of researchers. To Tom, my main collaborator over the year, thank you for your mentorship and comradeship, and making a collaboration over 5 time zones work as well as it did. To Fan, thank you for bringing me onto the Visual AI team and allowing me to gain so much exposure in Computer Vision. To other members of Visual AI, Maoran and Yuhao, I am thankful for your patience and generosity and for helping me navigate Linux commands, PyTorch, and training machine learning models when I first got here. To Nikita, thank you for asking hard questions throughout our collaboration on language detection and giving me the impetus to constantly refine and rethink my ideas. To Timur, you have been an incredible mentor in Econometrics and Urban Econ—the conversations we've had have been stimulating and enlightening. I am also sincerely grateful to the other members of Senseable and our collaborators in Stockholm for their valuable input over the year.

To Charles and Öz, who guided me through my early forays in research when I

was an undergraduate, and the numerous mentors I had in (the other) Cambridge—Michael, Dominique, Christian and so many more, thank you for your patience and mentorship, for giving me the best foundation I could have had entering graduate school.

As rewarding as my time working with Python and Overleaf on the 2nd floor of Building 9 has been, my time at MIT would have been incomplete without the friends I have met here. To the Singaporean MFin community who readily accepted me as a part-time MFin, the 9 months we spent together has been unforgettable—I will cherish the times we hopped around North America to whet our pandemic-deprived wanderlust, around campus for free food, and around my living room bouncing ping pong balls into red plastic cups.

To my classmates—Corwin, Michael, Edvard, Erlend, it has been a while since we had to grind out problem sets, but I truly cherish the times we spent in the Lewis conference room we basically claimed as our own, masquerading as student entrepreneurs in iHQ, and the occasional late nights in Hayden or the Stud. The CSE classes brought me into unfamiliar territory but the camaraderie has made this journey so much more manageable and fruitful.

Lastly, I have to thank my family and Xuan Yi—thank you for being an unconditional source of support over the years, for giving me the courage to dream and the tenacity to fulfill my dreams.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, I investigate how modern computer vision tools can characterize the visual quality and linguistic diversity of cities and leverage on these novel metrics to study the impact of new housing projects.

Cities are laden with visual clues. In my first visit to Flushing in New York City, I was surprised by the co-location of Chinese and Korean signs all around the neighborhood. I had been to Chinatowns and Koreatowns before but never a place that seemed like a mix of both. I later learned that Flushing had experienced large-scale immigration from East Asia in the late 20^{th} century and the presence of both Chinese and Korean in the streetscapes of Flushing today is emblematic of the cultural smorgasbord that is New York. While in New York, I also realized how much our experience in a city is shaped by its visual features. In midtown Manhattan, I was surrounded by towers after towers—awe-inspiring but also somewhat daunting. Although I am used to seeing skyscrapers in my hometown of Singapore, the ubiquity of trees lined along the streets of the "garden city" seemed to have a mellowing effect on the scale of city, unlike in Manhattan.

As Allan Jacobs points out, "[p]eople who live in cities ... take cues from their physical environments every day, knowingly or not, and they often base their actions

on those messages." [37] What we see and experience around us are not only clues about the condition of cities at present, but the changes in these visual features help us capture and understand the evolution of the city.

Previous work have looked at how to quantify these features in hopes of improving our understanding of cities [21, 22, 53]. However, these early efforts tend to take the form of field studies—either through making observations on the field or constructing features based on videos of streetscapes. Such efforts are laborious and not scalable to large urban areas such as an entire city. However, there are two key developments in the last decade that spell promise for this area of research:

In 2007, Google launched the first large-scale street-level online corpus, providing street-level photographs in thousands of cities—Google Street View (GSV). This kickstarted the explosion of street view imagery (SVI), with commercial competitors such as Bing and Baidu providing similar services. Today, it covers all 7 continents and includes temporal data for most major cities in the world. The sheer scale, both spatially and temporally, of GSV and of other forms of SVI provides an extensive dataset for researchers and policymakers to tap on to understand how cities evolve.

The second development, which also stems from the rise in big data, is the advent of deep learning in computer vision. In 2012, Krizhevsky et al. achieved a 10 percentage point decrease in classification error on the ImageNet Challenge[1] [41]. AlexNet (the model used by Krizhevsky et al.) utilized a deep convolutional neural network (DCNN) architecture and its impressive performance kickstarted the deep learning revolution in computer vision. In the last decade, we have seen the growth of deeper and better DCNN models [31, 35, 65], pushing the limits of what we can do with computer vision models. These new architectures provide powerful pre-trained models that can be further trained for more specific computer vision tasks such as semantic segmentation, scene classification and optical character recognition (OCR).

[1]The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale.

The growth of SVI and the advent of deep learning have also inspired the growth of computational social science, with researchers seeking insights from visual information in streetscapes. Salesses et al. and Dubey et al. trained models to predict how humans perceive their surroundings using GSV images [19,61]. Similarly, there has been work done to quantify greenery [8, 48], safety [73] and shade provision [47] by applying DCNNs to SVI. This thesis contributes to this new body of literature by using SVI to construct two sets of measures that characterizes the visual quality and linguistic diversity of streetscapes.

Beyond monitoring how measures of visual quality and linguistic diversity change spatiotemporally, I recognize that such changes need not happen spontaneously. Often, they are by-products of other changes. To that end, I investigate how these measures evolve in response to the construction of new housing projects. As part of the Senseable City Lab's collaboration with the City of Stockholm, I study how the construction of new housing in Stockholm affects the surrounding areas as measured by the novel metrics of visual quality and linguistic diversity.

1.1 Main Contributions

This thesis is divided into three parts. In **Chapter 2**, I review existing work seeking to characterize the visual quality of streetscapes and adopt a deep learning approach to construct informative features of streets. This work contributes to a growing body of literature that utilizes urban imagery to understand cities.

Given the poor performance of existing OCR tools in classifying languages present in a scene, I develop a language detection model in **Chapter 3**. The language detection model helps to characterize linguistic diversity in streetscapes and like our measures of visual quality adds another dimension to our understanding of cities.

In **Chapter 4**, I apply the novel metrics constructed in Chapters 2 and 3 to study the effects of new housing projects. While there is now a large body of literature in

using urban imagery to characterize cities, little has been done to study how these novel metrics respond to urban interventions. This thesis contributes to the literature by pushing the limits of how we can apply these novel metrics. Concurrently, this work also contributes to a strand of urban studies literature examining the spillover effects of new construction.

Chapter 2

Quantifying Visual Quality

Streets form a public space and how they look plays an important role in shaping how walkable they are, how safe people perceive them to be, and the general quality of living in the urban environment. Many developed cities such as London, Shanghai, Copenhagen have initiatives and guidelines to improve street design [12, 59, 71]. In fact, in the Stockholm City Plan, there is an overarching goal to provide "Good Public Spaces" [14] and under the Pedestrian Plan, the city has targeted for at least 85% of their residents feeling that the streetscape is attractive [13].

In line with these goals, it is important to develop a scalable process which can quantify visual changes in the streetscape, both spatiotemporally and in response to urban interventions. The traditional approach to quantify the visual quality of streetscapes involves field surveys and is limited to very small-scale empirical studies [70]. The advent of machine learning algorithms that can help quantify physical features of streetscapes spells hope for automating this process. Using the latest machine learning algorithms in scene classification and semantic segmentation, we construct measures of visual quality using physical features shortlisted by Ewing and Handy [22]. Recognizing that the visual quality of an urban area is not solely defined by what people see but also how people *feel* based on what they see, we train a model using the Place Pulse 2.0 dataset [19] that scores an image from 1 to 10 along 6

perceptual dimensions. We then apply these metrics to quantify the visual quality of Stockholm, and examine how these measures of visual quality vary spatially and their relationships with socioeconomic characteristics.

2.1 Literature Review

2.1.1 Visual Quality and Physical Features

What is considered a visually appealing street is subject to debate. In urban design literature, authors often cite intangible qualities that they hypothesize to intervene between physical features and human behavior. Such qualities are difficult to qualify and quantify. Experts may have an intuitive understanding of oft-used terms such as "enclosure" and "complexity" but these ideas are difficult to operationalize as there is no clear definition as to what they entail and how we can measure them. In an attempt to summarize and operationalize these subjective ideas of visual quality, Ewing and Handy studied how physical features relate to expert ratings along 5 urban design qualities—imageability, enclosure, human scale, transparency and complexity [22].

Ewing and Handy assembled a panel of 10 urban design and planning experts from professional practice and academia, with experience in both the US and Europe. They engaged these experts to study video clips of 48 commercial streets across 22 cities in the United States and provide ratings along 5 urban design qualities selected for their importance in urban design literature. Concurrently, the authors also analyzed each of the 48 video clips to construct more than 100 measures of physical features in the streets. The authors then regressed expert panel ratings against these physical features to establish objective features that are well-correlated with these subjective qualities (Table 2.1).

Although Ewing and Handy's extensive work has provided us with both urban design qualities and measurable features that we can use to characterize urban spaces,

Urban design quality	Significant physical features
Imageability	people (no.)
	proportion of historic buildings
	courtyards/plazas/parks (no.)
	outdoor dining (y/n)
	buildings with non-rectangular silhouettes (no.)
	noise level (rating)
	major landscape features (no.)
	buildings with identifiers (no.)
Enclosure	proportion street wall—same side
	proportion street wall—opposite side
	proportion sky across
	long sight lines (no.)
	proportion sky ahead
Human scale	long sight lines (no.)
	all street furniture and other street items (no.)
	proportion first floor with windows
	building height—same side
	small planters (no.)
	urban designer (y/n)
Transparency	proportion first floor with windows
	proportion active uses
	proportion street wall—same side
Complexity	people (no.)
	buildings (no.)
	dominant building colors (no.)
	accent colors (no.)
	outdoor dining (y/n)
	public art (no.)

Table 2.1: Significant physical features associated with each urban design quality. Reproduced from Ewing and Handy [22].

their manual approach in constructing these metrics is not scalable and would be ill-posed in characterizing the visual quality of large urban spaces. As Neckerman et al. who applied Ewing and Handy's work to auditing New York City note, there is a strong impetus to explore alternative methods of data collection to improve the time efficiency of quantifying the urban space [53]. In particular, they note that the preponderance of Google Street View (GSV) images, among other street view imagery (SVI) means that researchers no longer need to spend time collecting videos like Ewing and Handy. Furthermore, they also recognize the consistency of imaging techniques used in GSV, which ensures the spatial consistency of measures constructed from GSV images.

The explosion of big data has also contributed to the advent of deep learning. The rise of data-driven machine learning models that can automatically process images and extract high-level features spell promise for automating the laborious process of coding physical features in streetscapes. Indeed, Ye et al. studied how features easily identified by machine learning models can predict visual quality [72]. The authors suggest that measures produced by a semantic segmentation model—building frontage, greenery, sky view, pedestrian space, motorization and diversity—can capture the 5 qualities in Ewing and Handy's paper and they explore how these measures can predict a measure of visual quality. To construct the measure of visual quality, the authors invited 10 urban design experts to do pairwise comparisons of a representative sample of street view images in Shanghai, China before transforming the results of the pairwise comparison into scores using the Elo rating system. Like Ewing and Handy, Ye et al. study how pre-selected physical features can correlate with expert scoring. Their key contribution to literature is studying how existing machine learning algorithms can help automate the construction of the pre-selected physical features.

Although the authors' idea of leveraging existing machine learning algorithms to measure physical features in streetscapes marks a huge step in making the quantification of visual quality more scalable, the choice of measures Ye et al. used leaves more to be desired. Despite Ewing and Handy providing a long list of physical measures strongly correlated with each design quality, Ye et al. chose to only look at 5 features that can be measured with a semantic segmentation model.

They extract the percentage of pixels representing greenery, sky, buildings, pedestrian paths and pedestrians, and motorways and cars. Their measure of diversity is the percentage of pixels representing the rest of the design elements, thereby including features such as street lights and street furniture. Since all pixels in an image are represented in the model, there is no presupposition of what might be important ex ante. Furthermore, the authors go on to assert that diversity has the second highest relative importance even though it is really just a catch-all measure. Rather, there

are many more physical features and meaningful metrics that can be constructed with Ewing and Handy's work and Ye et al.'s machine learning-driven approach, a task we endeavor to complete in Section 2.2.

2.1.2 Perceptual Features

In both Ewing and Handy and Ye et al., the authors do not explicitly model *perceptual features*. Unlike physical features that are objective descriptors of a streetscape, perceptual features capture how people *feel* and what they *perceive* in the surroundings. Understanding how people feel in their surroundings is important given prior evidence of how perception of urban areas can affect socioeconomic outcomes [15, 20, 42].

Although urban design qualities draw from both physical and perceptual features in the streetscape, both Ewing and Handy and Ye et al. chose to only measure physical features under the implicit assumption that physical and perceptual features are well-correlated. Although Zhang et al. recognize that interesting correlations between physical and perceptual features exist, they also highlight that human perception goes beyond what we see and draws from our prior experiences in similar spaces [74]. After all, perception may have subtle or complex relationships with physical features [21]. Therefore, it is important to model perceptual features separately. Tang and Long measure perceptual features with ratings from an expert panel, and combine these measures with measures of physical features to construct an overall measure of visual quality [70]. In their study of temporal changes in Beijing *hutongs*, they engaged 4 experts with a background in urban design education to rate the willingness to stay in a space from 1 to 5, using criteria from Ewing and Clemente [21][1]. Like Ewing and Clemente and Ewing and Handy, such an approach is not scalable even if it is able to provide high quality information for a small area of study.

Rather, as Tang and Long note, previous work at MIT Media Lab by Dubey et

[1]Ewing and Clemente uses the same set of urban design qualities as Ewing and Handy, but replaces *complexity* with *tidiness*

al. [19] has investigated predicting perceptual features of a streetscape using deep convolutional neural networks. Dubey et al. construct a new crowdsourced dataset of urban appearance (Place Pulse 2.0)—using an online data collection platform, the authors show participants two street view images from the same city side-by-side and ask participants which place looks more safe, beautiful, depressing, lively, wealthy or boring. As of August 2020, the dataset comprises 1.55 million pairwise comparisons for 110,988 images from 56 cities in 28 countries across 6 continents, along the 6 perceptual dimensions. Place Pulse 2.0 builds on work in Salesses et al. [61] which introduced the original Place Pulse. Compared to Place Pulse, Place Pulse 2.0 has a much larger scale and higher visual diversity. Intuitively, increasing visual diversity would allow the model to generalize better and Dubey et al. show that holding the size of the dataset constant, increasing visual diversity improves the prediction accuracy of pairwise comparisons. Increasing the size of the dataset provides a further boost to prediction accuracy.

A key difference between Tang and Long's work and Dubey et al. is that Tang and Long relies on expert raters to rate the streets while Dubey et al. used ratings from the general public to construct a predictive model. There are academics who criticize the use of input from the general public in generating predictions vis-à-vis expert input on the grounds that the public does not necessarily know what is good design [70, 72]. However, since these measures of visual quality are subjective and the purpose of measuring these measures is ostensibly for the purpose of improving public satisfaction with public spaces, it makes sense for such measures to be labelled by the public. Furthermore, the participants in Dubey et al. hail from more than 150 countries, pointing to the diversity of responses—aggregating over such a large and diverse sample helps us to learn from the *wisdom of the crowd*.

In Dubey et al., the authors trained models with a Siamese architecture that takes in image pairs, extracts image features, and then predicts the winner of each pair. However, as Zhang et al. note, a model that can predict a score for a single sample (as opposed to a winner between an image pair) is more valuable for large-scale quan-

tification of perceptual features in an urban area [74]. Inspired by schedule methods used in Salesses et al. and Ordonez and Berg [56, 61], Zhang et al. constructed a process that translates pairwise comparisons into scores for individual images. In this process, an image is scored higher the more times it is chosen over another image, but the score of each image is also corrected according to the score of images it is compared to. We leverage on Zhang et al.'s scalable model and the latest data from Place Pulse 2.0 to construct perceptual features.

2.2 Building a Computational Model

2.2.1 Physical Features

Machine Learning Models

Like Ye et al., we use a semantic segmentation model to extract useful measures of physical features in streetscapes. Although Ye et al. and Tang and Long both use SegNet [5] which has an encoder network that is topologically identical to the 13 convolutional layers in the VGG-16 network [65], we apply an implementation with a ResNet-50 architecture[2] that yields better performance than SegNet for the ADE20K dataset[3]—41.26 Mean IoU[4] for our chosen implementation vs 21.64 for SegNet.

For a given image i defined by a 3D array $p_i \in \mathbb{R}^{l \times h \times c}$ where l is the length of the image, h the height of the image and c the number of color channels, the semantic segmentation model $SS(\cdot)$ returns a category (1 of 150 possible categories) for each

[2]Source code available at https://github.com/CSAILVision/semantic-segmentation-pytorch

[3]ADE20K is a common benchmark in semantic segmentation that consists of 150 object categories.

[4]Intersection over Union (IoU) is a common evaluation metric for semantic image segmentation. For each class, the IoU metric is defined as the number of true positives divided by the number of true positives, false positive and false negatives. Mean IoU is the mean of IoU across all classes.

of the $(l \times h)$ pixels.

$$SS(p_i) = \mathbf{S}_i = \begin{bmatrix} S_{i,1,1} & \dots & S_{i,l,1} \\ \vdots & \ddots & \vdots \\ S_{i,1,h} & \dots & S_{i,l,h} \end{bmatrix} \tag{2.1}$$

where $S_{i,j,k} \in [1, 150]$ is a positive integer representing the category of the pixel.[5] For most of the measures we construct, we are interested in the proportion of an image classified as a specific category i.e.

$$Prop_{i,m} = \frac{1}{l \times h} \sum_{j=1}^{l} \sum_{k=1}^{h} \mathbb{1}(S_{i,j,k} = m) \tag{2.2}$$

Since Ewing and Handy allude to physical features that are not necessarily objects present in image, but also the semantic interpretation of the scene altogether, we use the Places-CNN scene classifier [77] to provide other useful measures of streetscapes. Places-CNN is a scene classifier trained on the Places Database—a repository of 10 million scene photographs—to classify images into one of 365 scene semantic categories.[6] We use an implementation built on a ResNet-18 base architecture.[7]

Again, for a given image i defined by a 3D array $p_i \in \mathbb{R}^{l \times h \times c}$, the scene classifier $SC(\cdot)$ returns the vector of softmax probabilities for the 365 scene categories as an intermediate output.

$$\mathbf{C}_i = \begin{bmatrix} C_{i,1} & \dots & C_{i,365} \end{bmatrix} \tag{2.3}$$

where $\sum_{m=1}^{365} C_{i,m} = 1$. As a final output, the classifier returns

$$SC(p_i) = \mathbf{t}_i = \begin{bmatrix} t_{i,1} & \dots & t_{i,5} \end{bmatrix} \tag{2.4}$$

[5]Full list of ADE20K categories are presented in Appendix B
[6]Full list of Places scene categories are presented in Appendix B
[7]Source code available at: https://github.com/CSAILVision/places365

where $t_{i,k}$ is the k-th most likely scene category of an image i. In other words, $SC(\cdot)$ outputs the 5 most likely scene categories of an input image.

Constructing the Measures

Using the 5 urban design qualities in Ewing and Handy's work, we shortlist physical features from Table 2.1 that are also measurable in a scalable manner with modern machine learning tools. We discuss each of the 5 urban design qualities and the corresponding measures we construct below.

Imageability captures how distinct, recognizable or memorable a place is and is related to Cullen's idea of "sense of space" that entices people to enter the space [16, 21]. For example, Quincy Market in Boston may be considered a space with high imageability, as a hub of activity and liveliness, with strong architectural qualities which inspire people to gather. Following the features found to be significant in Ewing and Handy (Table 2.1), we measure the following:

1. $Prop_{i,13}$ i.e. proportion of pixels classified as people
2. $\sum_{k=1}^{5} \sum_{m \in M} \mathbb{1}(t_{i,k} = m)$ where M is the set of indices corresponding to the categories courtyard, plaza, outdoor diner and park i.e. number of scene categories within the top 5 predicted categories which are in the desired categories
3. $\sum_{m \in M} C_{i,m}$ where M is the set of indices corresponding to the categories courtyard, plaza, outdoor diner and park i.e. the combined probability of an image being classified as one of the desired categories

Although counting the number of people (or in this case, the proportion of pixels classified as people) from GSV images invites concerns about reliability, Ewing and Clemente find that pedestrian count in SVI (they used samples from Google, Bing and EveryScape) has almost perfect inter-rater agreement with manual audits on the

streets of New York City [21]. This finding has also been corroborated by Chen et al. through a validation exercise of much larger scale [10].

Enclosure captures how well streets and other public spaces are defined by vertical elements such as walls, buildings and trees [21]. Cullen sees enclosure as the "outdoor room" [16], which Jacobs argues help instills a sense of safety and memorability [38]. Following the features found to be significant in Ewing and Handy (Table 2.1), we measure the following:

1. $Prop_{i,1} + Prop_{i,2}$ i.e. proportion of pixels classified as wall or building
2. $1 - Prop_{i,3}$ i.e. proportion of pixels *not* classified as sky

Human scale captures how well physical elements in the streetscape coheres with the size and proportions of humans [21]. Although numerous urban designers and thinkers have offered specific definitions of what is considered human-scale, from the height of buildings to space for personal interaction to building widths, Ewing and Handy find that the presence of smaller-scale features correlates well with the degree of human-scale. Therefore, we measure the following:

1. $\sum_{m \in M} Prop_{i,m}$ where M is the set of indices corresponding to tree, grass, plant and flower i.e. proportion of pixels classified as street greenery
2. $\sum_{m \in M} Prop_{i,m}$ where M is the set of indices corresponding to sidewalk, table, chair, sofa, armchair, seat, desk and ottoman i.e. proportion of pixels classified as street furniture

Transparency captures the degree to which people can see or perceive human activity—it is more than what people can see in the streetscape but what they can *imagine* from the streetscape [21]. The most literal example of transparency would be windows but more subtle features such as signs that demonstrate specific uses can also add to transparency. Therefore, we measure:

1. $Prop_{i,9} + Prop_{i,44}$ i.e. proportion of pixels classified as windows or signboards

Complexity captures the visual richness of a place. Intuitively, high complexity makes streets more interesting to look at and Gehl suggests that this makes "the walking distance seem shorter" [23]. Although Ewing and Handy find that the presence of specific physical features contributes to visual complexity, these measures do not capture the idea of diversity that is implicit in the idea of visual complexity. Since we have already measured some of these physical features as part of our characterization of other urban design qualities—presence of people, outdoor dining, and buildings, we offer two alternative measures that better encapsulate the semantic meaning of complexity.

1. $\sum_{m=1}^{150} \mathbb{1}(Prop_{i,m} > 0)$ i.e. number of objects detected through semantic segmentation

2. $1 - \sum_{m=1}^{365} C_{i,m}^2$ i.e. $1-$ sum of squared softmax probabilities

In particular, the second metric used in complexity follows the Herfindahl-Hirschman Index commonly used in industrial organization

$$HHI = \sum_{i=1}^{k} P_i^2 \tag{2.5}$$

where $\sum_{i=1}^{k} P_i = 1$. The index is minimized when $P_i = 1/k \ \forall i \in [1, k]$. HHI is commonly used to measure market concentration and is minimized when the market is equally distributed among key players. In our case, we use $1-$ the sum of squared softmax probabilities so that a scene which is more complex (probabilities are almost equally distributed among the various categories) would have a higher score.

2.2.2 Perceptual Features

Translating Pairwise Comparisons to Individual Scores

We follow the method used in Zhang et al. [74] closely in our construction of perceptual features. For each image i, we define a positive rate (P_i) and a negative rate (N_i) as

$$P_i = \frac{p_i}{p_i + e_i + n_i} \tag{2.6}$$

$$N_i = \frac{n_i}{p_i + e_i + n_i} \tag{2.7}$$

where p_i (n_i) refers to the number of times i is (not) picked in a pairwise comparison and e_i the number of times i is considered equal to the other image. We then define a Q-score for each image i

$$Q_i = \frac{10}{3}\left(P_i + \frac{1}{p_i}\sum_{k_1=1}^{p_i} P_{k_1} - \frac{1}{n_i}\sum_{k_2=1}^{n_{k_2}} N_{k_2} + 1\right) \tag{2.8}$$

that ranges from 0 to 10.

Training Process

In Zhang et al., the authors assigned a binary label to each image based on its Q-score crossing a certain threshold and formulated the prediction problem as a binary classification task. Although this method was implemented to create a gap between "positive" and "negative" samples by removing "noisy" data in the center, there is significant data loss and loss in granularity.

Therefore, we construe the problem as a multi-label classification task with 10 labels (representing integers from 1 to 10), following the approach taken in [73]. To translate the Q-scores into labels from 1 to 10, we take the following steps:

1. Standardize the Q-scores along each perceptual dimension to get zero mean and

Hyperparameter	Choice
Base Architecture	DenseNet-121
Learning Rate	$0.000001, 0.000005, 0.00001, 0.00005$
Momentum	0.9
Weight Decay	$10^{-4}, 10^{-3}$
Batch Size	130
Epochs	150
Loss Function	Cross-entropy
Optimizer	Stochastic gradient descent
Weights	[5, 3, 2, 1, 1, 1, 1, 2, 3, 5], [4, 3, 2, 1, 1, 1, 1, 2, 3, 4], [4, 3, 2, 2, 1, 1, 2, 2, 3, 4], [3, 3, 2, 2, 1, 1, 2, 2, 3, 3], [5, 4, 2, 1, 1, 1, 1, 2, 4, 5]
Image size	256×256

Table 2.2: Hyperparameters for training perceptual features classification model

unit variance using `StandardScaler` in the Python `sklearn` package

2. Divide the data into 10 bins of equal-width

3. Store the center value of each bin: yields $\mathbf{m}_f \in \mathbb{R}^{10}$ for each perceptual feature f

4. Assign a label to each image based on which bin it is in

The base code for training and testing follows a modified version of the `PyTorch` implementation from [77] that supports multi-label classification.[8] We experiment with the learning rate and weights allocated to each of the 10 labels.

Since the model uses a softmax classifier, an intermediate output in the multi-label classification is the softmax probabilities over each each of the 10 possible labels. Using the information of center values for each bin $\mathbf{m}_f$, we can reconstruct fine-grained scores:

1. Compute the normalized score $s_{i,f} = \boldsymbol{\sigma}'_{i,f}\mathbf{m}_f$ where $\boldsymbol{\sigma}'_{i,f}$ is the vector of softmax probabilities of image i for feature f

[8]Modifications made by Fan Zhang from MIT Senseable City Lab.

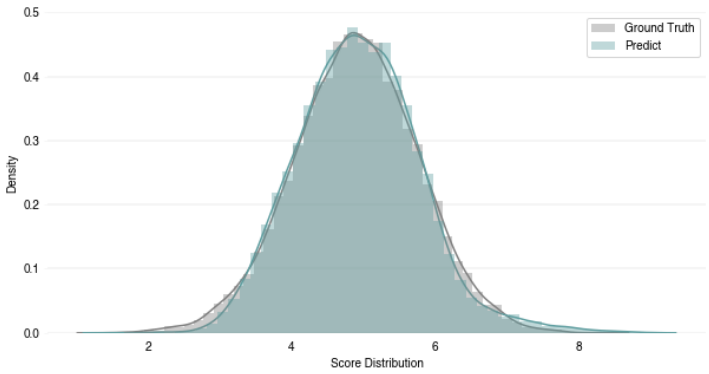

Figure 2-1: Distribution of predicted scores along all 6 perceptual dimensions corresponds well with ground truth; model yields a mean absolute percentage error of 0.168

2. Denormalize the score using `StandardScaler` with the mean and variance of the Q-scores for the relevant feature

Another point of deviation from Zhang et al. [74] is the training dataset. The Place Pulse 2.0 dataset has continued expanding since the publication of [74], with a 50% increase in data available. The larger dataset indeed leads to improved performance, yielding a mean absolute percentage error of 0.168, compared to 0.183 when trained with the smaller dataset. Figure 2-1 shows how the distribution of the predicted scores compares with the distribution of the ground truth.

2.3 Applications in Stockholm

For the twin purposes of validation and exploratory analysis, we apply the models to our case study of Stockholm. Using images of Stockholm extracted from GSV,[9] we construct fine-grained measures of visual quality and compare them with other neighborhood characteristics.

[9]Details of data extraction are outlined in Appendix A.

	Population Density	Median Income
Imageability (People)	0.420	0.302
Imageability (SC, Top 5)	-0.105	-0.104
Imageability (SC, Prob)	-0.111	-0.133
Enclosure (Wall/Building)	0.731	0.417
Enclosure (Sky)	0.639	0.349
Human Scale (Greenery)	0.178	0.198
Human Scale (Furniture)	-0.114	-0.115
Transparency	0.0473	0.0497
Complexity (Objects)	-0.00526	0.0547
Complexity (Herfindahl-Hirschman)	-0.119	-0.137

Table 2.3: Correlation of physical measures of visual quality with neighborhood characteristics in Stockholm

2.3.1 Physical Features

As a form of validation, we look at how the physical measures of visual quality correspond with neighborhood characteristics. Following the work of Neckerman et al. in New York City, we look at how visual quality correlates with population density and median income. We use data of population density from 2010 to 2021 and median income from 2011 to 2021 from Statistics Sweden [68] . The correlation coefficients are presented in Table 2.3. Figure 2-2 presents choropleth maps of population density, median income and all 10 physical measures of visual quality in Stockholm at the DeSO (Demographic Statistical Areas) level in the period 2020-2021.

Population Density

We find that population density has a moderate positive correlation with imageability as measured by the proportion of pixels classified as people. This is intuitive since higher population density should correspond to higher pedestrian traffic. However, population density has a weak negative correlation with imageability as measured by the outputs of the scene classifier. This is counterintuitive since we expect the likelihood of seeing courtyards, plazas, outdoor diners, parks to be higher in more densely populated areas, which suggests that imageability measures constructed with

the scene classifier may have limited usefulness.

Intuitively, population density should correspond with a denser built-up area that is livelier. This corresponds to higher scores on enclosure and transparency. Indeed, we find strong positive correlation between population density and both measures of enclosure and weak positive correlation between population density and our measure of transparency.

Neckerman et al. argue that population density may have a negative association with human scale due to the denser and possibly taller built-up structures [53]. However, Ewing and Handy note that street elements such as trees, or features that demonstrate street-level activity such as furniture can moderate the scale of tall buildings, citing Times Square and Rockefeller Center in New York City [22]. We find that population density has a weak positive correlation with human scale as measured by street greenery, and a weak negative correlation with human scale as measured by street furniture. The overall effect is ambiguous, as we expect, given the lack of consensus in literature.

We expect population density to have a positive association with complexity since it is more likely for denser areas to have more street-level activity, buildings with mixed uses and visual features that contribute to overall visual richness. That said, we find that complexity as measured by the number of objects is uncorrelated with population density while complexity as measured by the Herfindahl-Hirschman index has a negative association. The latter reinforces the notion that second-order metrics from the scene classifier may have limited usefulness in characterizing urban streetscapes, likely because there is a lot more underlying noise in its predictions than in a semantic segmentation model.

Median Income

Like Neckerman et al., we hypothesize that urban areas with higher median income should have higher scores across the board since higher visual quality should be priced into rent and housing prices, which would correspond to a more affluent populace. Table 2.3 shows that this is mostly true, except for imageability as measured by the scene classifier, human scale as measured by street furniture and complexity as measured by the Herfindahl-Hirschman index.

2.3.2 Perceptual Features

We present choropleth maps of perceptual features in Stockholm in 2020-21 in Figure 2-3.

In general, we notice a small spread in these indices throughout most of the city, with only the boroughs of Hässelby-Vällingby and Spånga-Tensta standing out as being less beautiful, more depressing, less safe, less wealthy, but marginally less boring and livelier.

Although it is unclear why these boroughs are considered to be livelier, their scores for the other perceptual dimensions agree with our understanding of Stockholm. The two boroughs are centered on large public housing projects—Vällingby, the first ABC town[10], Hässelby gård and Tensta, both part of the Million Programme.[11] These projects are known to be inexpensive, but architecturally dull [29]. Furthermore, both boroughs are considered to be among the most vulnerable areas in Sweden, with higher levels of crime and social exclusion [4]. Therefore, it is no surprise that their perceptual scores stand out from the rest of Stockholm.

[10]Loosely translated as a labor housing center, acting as a self-contained city, providing both employment and housing.

[11]The Million Programme is a government initiative to build 1 million new dwellings in Sweden that ran between 1965 and 1974.

	Beautiful	Boring	Depressing	Lively	Safety	Wealthy
Imageability (People)	0.124	0.0840	-0.0881	-0.115	0.0609	0.0592
Imageability (SC, Top 5)	-0.373	-0.352	0.156	0.525	-0.0942	-0.123
Imageability (SC, Prob)	-0.457	-0.425	0.223	0.601	-0.152	-0.171
Enclosure (Wall/Building)	0.234	0.116	-0.160	-0.202	0.119	-0.126
Enclosure (Sky)	0.214	0.128	-0.150	-0.206	0.105	0.110
Human Scale (Greenery)	0.923	0.307	-0.738	-0.664	0.600	0.676
Human Scale (Furniture)	-0.538	-0.401	0.345	0.561	-0.238	-0.304
Transparency	0.0349	0.0989	0.0458	-0.146	-0.0817	-0.0874
Complexity (Objects)	0.00813	0.111	0.0462	-0.0968	-0.0691	-0.0385
Complexity (HH)	-0.497	-0.318	0.312	0.511	-0.316	-0.274

Table 2.4: Correlation of physical measures of visual quality with perceptual features in Stockholm

2.3.3 Relationships between Physical and Perceptual Features

Although we construe visual quality as the synthesis of physical features and perceptual features, we recognize that they are not mutually exclusive. Rather, Zhang et al. and Ewing and Clemente both recognize that what we *see* and what we *feel* are inextricably linked, even if they do not fully overlap [21, 74]. In Table 2.4, we present the correlation coefficients between physical features and perceptual features.

We see that human scale, as measured by the presence of street greenery has strong positive correlation with perceived beauty, safety and wealth but is negatively correlated with how depressing a place feels. The effects of greenery has been well documented in urban studies literature, with Ashihara arguing that urban greenery offers a sense of peacefulness and quietness [2] and Knez showing in a field study in Gothenburg, Sweden that urban greenery leads to improved well-being [39].

We also notice some interesting relationships between the physical measures and liveliness. Imageability, as measured by the scene classifier, and human scale, as measured by street furniture, has a moderate-to-strong positive correlation with liveliness. This is intuitive since these measures of imageability and (to some extent) human scale proxy for human activity at the street level, which contributes to the liveliness of an urban area. Complexity, as measured by the Herfindahl-Hirschman Index, also has moderate-to-strong positive correlation demonstrating how mixed uses

may relate with the liveliness of an urban area. Interestingly, human scale, as measured by street greenery exhibits moderate negative correlation with liveliness. Even though greenery contributes to how scenic or visually appealing a streetscape is, it does not necessarily correlate with human activity. For example, Times Square in New York City, though extremely vibrant, is almost devoid of trees. Rather, this finding highlights the importance of constructing measures along different perceptual dimensions to capture a multi-faceted view of cities.

2.4 Concluding Remarks

In this chapter, we develop a scalable process to quantify the visual quality of a streetscape. Building on the work done by Ewing and Handy and Dubey et al., we characterize a streetscape along 5 urban design qualities (through measures of physical features) and 6 perceptual dimensions. We leverage on the abundance of SVI offered by GSV and the strong performance of deep learning approaches in computer vision to automate the process of extracting visual features from SVI and scoring a streetscape along the different dimensions of visual quality. We apply the model to Stockholm, Sweden and find that most of our measures of visual quality exhibit intuitive relationships with socioeconomic characteristics and with one another. In Chapter 4, we will further explore how informative measures of visual quality are in the context of evaluating the effects of new housing projects.

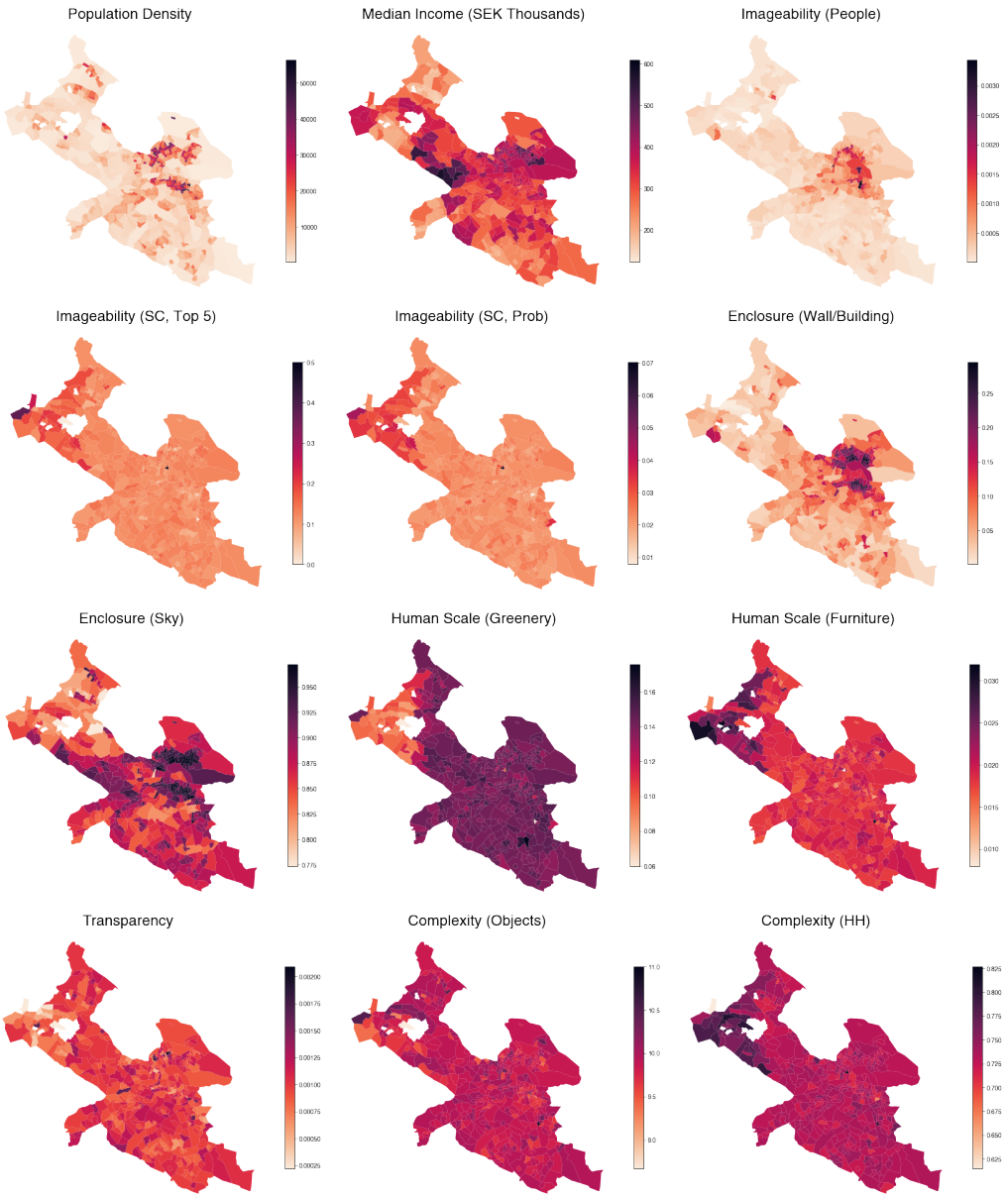

Figure 2-2: Choropleth maps of population density, median income and all 10 physical measures of visual quality in Stockholm at the DeSO level in the period 2020-2021

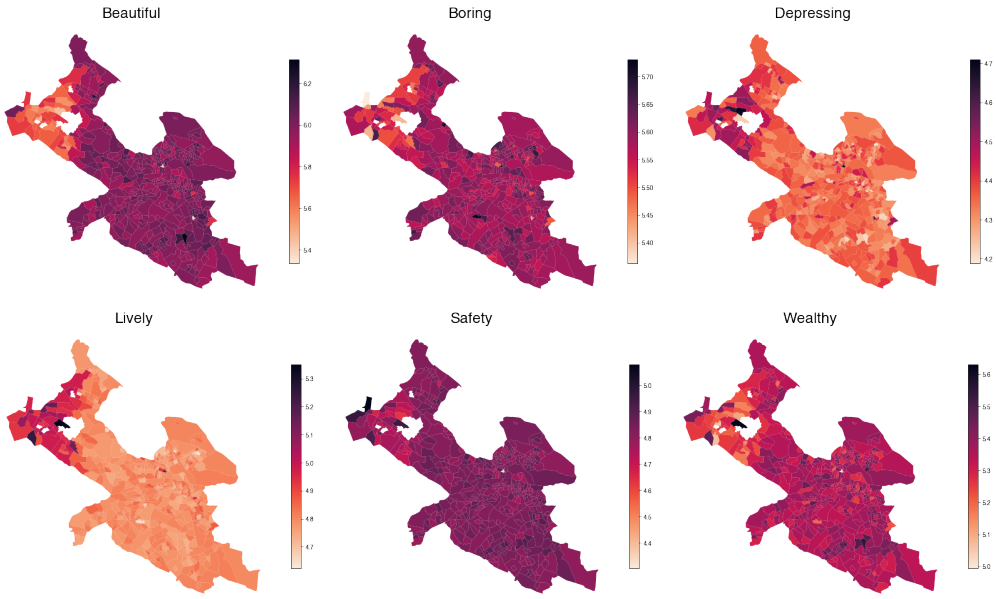

Figure 2-3: Choropleth maps of perceptual features in Stockholm at the DeSO level in the period 2020-2021

Chapter 3

Quantifying Linguistic Diversity

Any visitor to New York City would hardly find it difficult to locate an ethnic enclave. Chinatown, Little Italy, Koreatown, Little Egypt are all easily identifiable on a map, and anyone who unknowingly walks into these places would quickly figure out where they are from street signs, names of restaurants, advertisements and other language clues (Figure 3-1). Yet, not all enclaves have self-explanatory names like those in a major city. Nothing about the names Skärholmen (Stockholm) or Edgware Road (London) suggests that they have a large Arabic-speaking population, but visual clues from the languages that appear on these streets suggest otherwise.

Although census data may provide fine-grained information on the ethnolinguistic distribution of residents across census tracts, visual information provides a different perspective, giving us clues not only about the people who live there but also about people who work in and frequent the area. When combined with census information, visual clues may give us an idea of the level of social integration within an area—whether linguistic diversity is commensurate with the ethnic diversity of the urban area. From an urban policy standpoint, it is important to be cognizant of the ethnolinguistic make-up and linguistic diversity of urban regions and monitor how these metrics evolve spatiotemporally. Therefore, we seek to use modern computer vision tools to quantify the linguistic diversity of streetscapes.

Figure 3-1: Streetscape of Chinatown, Little Italy, Little Egypt, Koreatown from top left in clockwise direction.
Sources: "Chinatown, NYC" by nmadhu2k3, "Little Italy, NYC" by RobertFrancis and "Koreatown NYC" by Chun's Pictures are licensed under CC BY 2.0. Image of Little Egypt is taken from GSV.

This study is especially interesting in Stockholm, and Sweden more generally. Sweden has been characterized by a common language, religion, and political history for most of its history [17]. Amidst waves of immigration, most recently the large influx of Arabic-speaking people in the aftermath of the Arab Spring, it is interesting to see how native Swedes cope with living with a growing population of people who look and speak differently and how this manifests in the linguistic diversity on the streets.

Current tools that can be used for language classification are generally developed for optical character recognition (OCR) and the performance is mixed. Therefore, we take a different approach—rather than trying to identify every character in a scene, we seek to detect the presence of specific languages. Given that we are only interested in the prevalence of languages in a city rather than the actual words used, we can sacrifice the complexity of the problem to achieve better accuracy.

Since we aim to study linguistic diversity in streetscapes in Stockholm, we construct a novel multilingual dataset containing Google Street View (GSV) images from

7 cities, covering the presence of 4 languages—English, Swedish, Arabic and Chinese. We train a binary classification network over 4 independent tasks using a pre-trained DenseNet, and our best model achieves a test accuracy of 80.8% with a corresponding F1 score of 79.8%.

We apply this model to GSV images in Stockholm to further test the model's usefulness. We find that the spatial distribution of language concentration in Stockholm agrees with intuition. Although somewhat counterintuitive, we find that language mix does not correspond well with population mix,[1] instead providing a different perspective of Stockholm. We argue that in a culturally homogeneous society such as Sweden, minorities may not necessarily feel comfortable expressing themselves in their languages, which is reflected in the chasm between population and linguistic mix.

3.1 Literature Review

3.1.1 Linguistic Diversity of Streetscapes

The study of linguistic diversity in streetscapes is more commonly known as the study of linguistic landscapes in sociolinguistic literature. This body of literature is motivated by the abundance of linguistic features littered across streetscapes—public road signs, advertising billboards, street names, place names, commercial shop signs, street art, and public signs on government buildings [43]. As Gorter eloquently puts it: "[s]igns are everywhere, they permeate our daily life, and they can give us a sense of place" [25]. However, as much as signs influence us and our behaviors, they are also artifacts of individual and social preferences [54]. After all, the languages used on a sign is an outcome of deliberate choices made by political or economic actors,

[1]Since fine-grained data on ethnic/racial distribution is not publicly available, we construct an entropy measure with broad categories used in Statistics Sweden's publicly available data (Section 3.5.2).

which are in turn influenced by both economic interests and the social preferences of the populace. In a case study of Donostia, Spain and Ljouwert, the Netherlands, Onofri et al. find that the choice of languages in a sign is strongly influenced by the type of establishment associated with the sign. For instance, the use of English is strongly positively correlated with an establishment being an international chain. In contrast, the use of Frisian in Ljouwert is strongly correlated with the establishment being a shop or an official building [54].

As much as we can learn something about individual and social preferences from the languages we see in street signs, we may also learn something from those we do *not* see. In a field study of Oslo, Norway, Opsahl finds scarcely any presence of the Polish language, despite the large number of Polish immigrants in Norway [55]. Although we often see cities as platforms for languages to "manifest their vitality as well as their visibility" [6], the counterintuitive "invisibility" of Polish makes one question if individuals can indeed express themselves in their preferred tongue in Oslo [55]. And perhaps this is a valid question not just in Norway but also in Sweden, with Leinonen and Toivanen arguing that the Nordic nations have an identity that builds on cultural, religious, and linguistic homogeneity [44].

Even though studying linguistic landscapes is interesting, research in this area often relies on extensive field work in small communities. Such an approach offers deep insights into specific urban areas, but these insights may not necessarily generalize to the larger urban environment. Given the abundance of street view imagery (SVI) offered by GSV and similar services, there is a nascent field that seeks to automate the process of characterizing linguistic landscapes. In particular, Hong conducted a proof-of-concept study of a small Chinese community in Seoul, South Korea using Google Vision API tools [33]. Albeit their approach sets a new direction for studying linguistic landscapes, they find that many word sequences in their samples were unrecognized by the algorithm. To advance this literature further, we also take a big-data approach in constructing fine-grained maps of linguistic diversity in Stockholm.

3.1.2 Scene Text Recognition

The problem of scene text recognition is part of a larger field of image-based sequence recognition. The fundamental motivation in this field of research is to extract as much high-quality information from images as possible. Therefore, current tools that can be used for language classification are designed to not only detect if a language is present in an image but also output what the specific word sequences are.

EasyOCR[2] is a popular open source ready-to-use tool for scene text recognition. The tool uses a convolutional recurrent neural network (CRNN) architecture [64] that integrates feature extraction, sequence modeling and transcription into a unified framework to identify text in a scene. EasyOCR supports over 80 languages and has reasonable performance for English. However, it has a distinctly poorer performance for identifying other languages (see Section 3.4 for a comparison with our model).

There is also a host of commercial OCR tools that spells more promise, although it is unclear how these tools function. We experiment with Google OCR, which does a reasonably good job of recognizing specific characters when it detects the presence of any word sequence. However, it has a low recall, which corroborates with the findings of Hong [33]. This translates to an average accuracy and an F1 score poorer than EasyOCR (see Section 3.4). The mixed performance of state-of-the-art OCR tools motivates adopting a different paradigm in scene text recognition.

3.1.3 Classification with CNNs

Therefore, we seek to tackle an easier task—identifying the presence of a language in a scene. Like how OCR tools use some form of convolutional neural network (CNN) to extract features, our language detection model is also built on a pre-trained CNN.

Common CNNs used in extracting features from a scene include VGG [65] and

[2]Source code available at: https://github.com/JaidedAI/EasyOCR

ResNet [31]. While VGG offers a more parsimonious network (19 layers), the increased complexity of ResNet (>100 layers) accords more degrees of freedom for better feature representation. Against the backdrop of deeper and deeper networks and the associated problem of vanishing gradients, the Dense Convolutional Network (DenseNet) [35] offers a different modeling paradigm. In each layer of a DenseNet, the feature maps of all preceding layers are used as inputs, thereby creating a *denser* representation structure with fewer layers, striking a balance between model parsimony and representation space. DenseNets have performed well in image and scene classification tasks [75,76], and we use a pre-trained DenseNet in our implementation.

3.2 Data

3.2.1 Synthetic Data

Given the labor cost of manual labeling, we automate the process of data generation by using SynthText,[3] a tool for generating text onto given background images. Generating synthetic data is a common technique used in scene text recognition [30, 49] due to limited authentic data. To synthesize images with text, Gupta et al. identify regions with sufficient continuities using segmentation data and transform the text to be placed on images using depth data [28].

Although SynthText was originally developed only for English, we made amendments to the original code to provide support for generating Swedish, Arabic and Chinese text. We do this by introducing a multilingual corpus and fonts that support the three other languages.

Given that we can change the language of the text generated while controlling for the background image, the data generated may skew training towards focusing on the text rather than irrelevant visual features. In Figure 3-2, we provide examples of 4

[3]Source code available at: https://github.com/ankush-me/SynthText

Figure 3-2: Examples of synthetic images generated with the same background but with different languages—Arabic, Chinese, Swedish and English, from top left in clockwise order.

images with exactly the same background, each with text of a different language. We provide statistics of the synthetic dataset in Table 3.1.

3.2.2 Google Street View Images

Although there are benefits in using synthetic data, we recognize that visual domain adaptation is difficult [62]. Since we ultimately want to apply our text detection tool to real scenes, it is important for the model also to be trained on real streetscapes. Therefore, we source real data from GSV and manually label them with `pigeon`,[4] an open source labeling tool. With the intention of applying this tool to Stockholm, we scrape images from densely populated areas with a good amount of text in the scene. Our dataset comprises images from cities where the target languages are dominant—Stockholm for Swedish, Ramallah, Bethlehem and Beirut for Arabic and Hong Kong for Chinese. To avoid overfitting on these cities where the target languages are highly prevalent, we also include images from London and New York City,

[4]Source code available at: https://github.com/agermanidis/pigeon

Language	Synthetic	Real
English	1028	6105
Swedish	1028	1175
Arabic	1028	1434
Chinese	1028	1165
None	1028	6049
Total	5140	15928

Table 3.1: Size of dataset for each language. The synthetic data for each of the 4 languages is generated from the same 1028 background images. After finding good performance with training solely on real data, we scaled up manual labeling of real data, thereby leading to a much larger dataset of real data.

both global metropolis where minority languages feature in a less prominent manner. Therefore, we have a dataset with good coverage of the four target languages, with variations in how prominently they are featured and the architectural styles of the scenes they are featured in. This ensures that the resultant model is generalizable to our case study of Stockholm.

To obtain GSV images, we begin by generating sampling points along the road network in areas of interest at 50-meter intervals using OpenStreetMap. We then make API requests from GSV using these sampled coordinates and the following parameters—90° field-of-vision, 0° pitch, 50m radius. For each set of coordinates, we obtain images at compass headings of 0°, 90°, 180° and 270°, thereby capturing the full panorama at each point. The summary statistics of data obtained from each city is presented in Table 3.2.2.

We first approached the study with the hypothesis that pre-training with synthetic data would allow for better overall performance. Therefore, we generated synthetic data and labeled a small amount of real data manually. After finding better performance with training solely on real data vis-à-vis pre-training with synthetic data, we scaled up manual labeling of real data, thereby leading to a much larger dataset of real data (Table 3.1).

City	En	Sv	Ar	Cn	None	Total
London	2836	0	103	0	2543	5383
New York City	355	0	29	0	302	660
Stockholm	590	1175	2	15	1195	2633
Ramallah	873	0	1048	0	988	2214
Bethlehem	188	0	217	0	240	522
Beirut	48	0	35	0	59	127
Hong Kong	1215	0	0	1150	722	2034
Total	6105	1175	1434	1165	6049	13573

Table 3.2: Size of dataset by city

3.2.3 Dealing with Imbalanced Data

Since the language detection model is construed as a binary classification model with 4 parallel tasks, we will have many more negative examples than positive examples. For example, if we train the classifier solely on synthetic data where each image only has one language present, each classifier will have 1028 positive examples and 4112 negative examples. Therefore, we downsample the data randomly before training so that the number of positive examples and negative examples are equivalent.

3.3 Building a Computational Model

3.3.1 Hyperparameters

We train the models using a NVIDIA GeForce RTX 2080 Ti GPU with 11GB memory. We adopt a train-validation-test split of 70-15-15 and experiment with different learning rates before using the model with the best validation accuracy on our test set. The remaining hyperparameters are listed in Table 3.3.

The base code for training and testing follows a modified version of the `PyTorch` implementation from [77] that supports multi-label classification.[5]

[5]Modifications made by Fan Zhang from MIT Senseable City Lab.

Hyperparameter	Choice
Base Architecture	DenseNet-121
Learning Rate	$0.0005, 0.001, 0.002$
Momentum	0.9
Weight Decay	10^{-4}
Batch Size	100
Epochs	150
Loss Function	Cross-entropy
Optimizer	Stochastic gradient descent
Image size	256×256

Table 3.3: Hyperparameters

3.3.2 Training Process

We first train the model purely on the synthetic dataset, before training the best performing model on the real dataset for another 150 epochs. We also train the model with the real dataset from scratch.[6]

3.3.3 Second-order Metrics

The model predicts if a language is present in an image (outputs 1) or not (outputs 0), lending itself to a measure of language concentration in an urban area as defined as:

$$P_{x,\ell} = \frac{1}{|x|} \sum_{i \in \mathcal{S}(x)} \mathbb{1}(\text{language } \ell \text{ is in image } i) \tag{3.1}$$

where $\mathcal{S}(x)$ is the set of images in an urban area x. To quantify the linguistic diversity in an urban area, we use the concept of Shannon entropy from statistical physics that is commonly applied to capture the notion of diversity [1, 51, 67].

$$E_x = \sum_{l} -P_{x,\ell} \log(P_{x,\ell}) \tag{3.2}$$

[6] i.e. from pre-trained DenseNet-121 parameters

3.3.4 Evaluation Metrics

To evaluate the performance of our model, we look at both classification accuracy and the F1 score, which is defined as such:

$$\text{Precision}_\ell = \frac{TP_\ell}{TP_\ell + FP_\ell} \tag{3.3}$$

$$\text{Recall}_\ell = \frac{TP_\ell}{TP_\ell + FN_\ell} \tag{3.4}$$

$$\text{F1}_\ell = 2 \cdot \frac{\text{Precision}_\ell \cdot \text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell} \tag{3.5}$$

where TP_ℓ (FP_ℓ) is the number of positive examples that are (in)correctly classified for language ℓ while FN_ℓ is the number of negative examples that are incorrectly classified for language ℓ. Although classification accuracy is easily interpretable, it is not robust to data imbalance. Since we have a larger proportion of negative examples than positive examples for our test dataset, we are concerned about misleading test statistics (for example, a model may achieve high test accuracy by classifying everything as negative). Tracking recall and the F1 score provides us with another perspective of the model's performance.

3.4 Results

3.4.1 OCR vs Our Method

The test accuracy of the different models is presented in Table 3.4. We use our own test dataset, comprising 2964 images scraped from GSV, to evaluate the performance of our model and compare them to those of ready-to-use OCR tools. We find that both OCR tools have around 60-70% accuracy across the 4 languages.

From the precision and recall of the different models (Tables 3.5 and 3.6), we find that the OCR tools are generally quite conservative, as precision is generally

much higher than recall across the 4 languages. This is especially the case for Google OCR, which has almost no false positives, at the expense of a large number of false negatives.

On the other hand, we find that our language detection model works well for all 4 languages, when trained with real data. The overall accuracy of both models trained with real data is higher than that of both OCR tools and crucially, the gap between precision and recall for our models is small. In fact, for the models trained with real data, the gap between precision and recall for each classifier only has a maximum of 11.5 percentage points, much smaller than those of the 2 OCR tools. Consequently, the F1 scores of the models trained with real data are also higher than those of the OCR tools (Table 3.7).

Intuitively, the superior performance of our model in comparison with OCR tools arises from the fact that it is not trained to care about the specific characters. In an OCR tool, the model is trained to identify word sequences and only yields any output if the model has a sufficiently high confidence that a particular word sequence is present in the image. The language identified in the word sequence is only a byproduct and depends on a word sequence being identified first. However, identifying a language does not necessarily require identifying the specific words first, even for humans, particularly when the languages of concern are distinct from one another. In our specific use case, we note that Arabic is a cursive language where most of the characters in one word are connected, while written Chinese comprises pictographs that are of roughly equal size. These characteristics make Arabic and Chinese distinct from the Latin script and from each other. Even between English and Swedish, there are clear visual distinctions, with the use of accents and much higher frequency of long compound words in Swedish (e.g. Centralstationen (Swedish) vs Central Station (English)). Not focusing on specific characters allows the model more degrees of freedom to focus on distinct linguistic features, and to output a positive result, even when it is not clear what the specific words are (e.g. in scenes where the words are small or skewed).

Model	En	Sv	Ar	Cn	Total
EasyOCR	75.2	67.0	68.6	64.0	71.9
Google OCR	71.6	67.6	74.7	62.6	70.5
Synthetic	55.3	53.4	48.1	53.1	53.9
Synth + Real	71.6	82.4	**88.6**	88.9	77.2
Real	**76.4**	**85.5**	88.1	**91.1**	**80.8**

Table 3.4: Test accuracy of EasyOCR, Google OCR and our models. For each training paradigm, we only include the test accuracy of the model with the highest total validation accuracy. Highest test accuracy bolded.

3.4.2 Training with Synthetic vs Real Data

The performance of the model trained solely on the synthetic data is poor, with the test accuracy being about as good as a random guess. This points to the limitation of transfer learning—the underlying distributions of the synthetic data and real data are different and a model trained solely on synthetic data may not generalize well to real data.

However, once we train the model further with real data, we find improved performance and the test accuracy for languages other than English is much higher than that of the OCR tools. We also trained the model from scratch with real data and this model achieves better performance than OCR tools across all 4 languages, with a total test accuracy of 80.8%. Although we hoped that pre-training with synthetic data would teach the model to focus on the text rather than irrelevant features and therefore lead to better performance than training from scratch, it is likely that the poor generalization of the synthetic data made the parameters learned from training with synthetic data a distraction rather than an aid. The strong performance of training from scratch suggests that it may not be necessary to use synthetic data to train a language detection model.

Model	En	Sv	Ar	Cn
EasyOCR	82.2	65.7	68.7	66.1
Google OCR	71.2	94.1	100.0	98.0
Synthetic	68.7	42.9	47.2	60.3
Synth + Real	69.2	76.2	86.7	87.6
Real	79.2	79.0	85.6	90.2

Table 3.5: Precision of EasyOCR, Google OCR and our models. For each training paradigm, we only include the precision of the model with the highest total validation accuracy.

Model	En	Sv	Ar	Cn
EasyOCR	63.3	56.6	65.1	60.9
Google OCR	71.2	30.2	47.8	27.4
Synthetic	18.9	9.4	57.4	24.6
Synth + Real	76.4	88.7	90.4	91.1
Real	70.7	92.5	90.9	92.7

Table 3.6: Recall of EasyOCR, Google OCR and our models. For each training paradigm, we only include the recall of the model with the highest total validation accuracy.

Model	En	Sv	Ar	Cn	Total
EasyOCR	71.5	60.8	66.9	63.4	68.4
Google OCR	71.2	45.7	64.7	42.8	63.4
Synthetic	29.6	15.4	51.8	34.9	30.8
Synth + Real	72.6	82.0	**88.5**	89.3	77.8
Real	**74.7**	**85.2**	88.2	**91.4**	**79.8**

Table 3.7: F1 score of EasyOCR, Google OCR and our models. For each training paradigm, we only include the F1 score of the model with the highest total validation accuracy. Highest F1 score bolded.

3.4.3 Visual Interpretability

Although our model performs well, we want to be sure that it is looking at the right features instead of picking up spurious correlations. Therefore, we use a gradient-weighted class activation map (Grad-CAM) to translate gradient information of each of the 4 classifiers flowing into the final convolutional layer onto a heatmap.[7] In Figure 3-3, we present the heatmaps produced for a sample of correctly classified images in the test dataset. In general, we find that in correctly classified examples, the model focuses on the right things—store signs, road markings. In incorrectly classified examples (Figure 3-4), the model might still be looking at the right things but fails to detect a language likely because the word sequences are too small or indistinct.

We also provide the Grad-CAM of correctly classified multilingual scenes (Figure 3-5 provides five examples—English and Swedish, English and Arabic, English and Chinese, Swedish and Arabic, and Swedish and Chinese). In general, we find that the pair of relevant classifiers tends to focus on the same spot, likely since multilingual text are often co-located with English. In the English and Arabic example, where the English and Arabic text are not co-located, we see that the English classifier focuses on the part containing English text and the Arabic classifier focuses on the part containing Arabic text, which suggests that each classifier focuses on the correct target language.

[7]Modified implementation of: https://github.com/eclique/pytorch-Grad-CAM

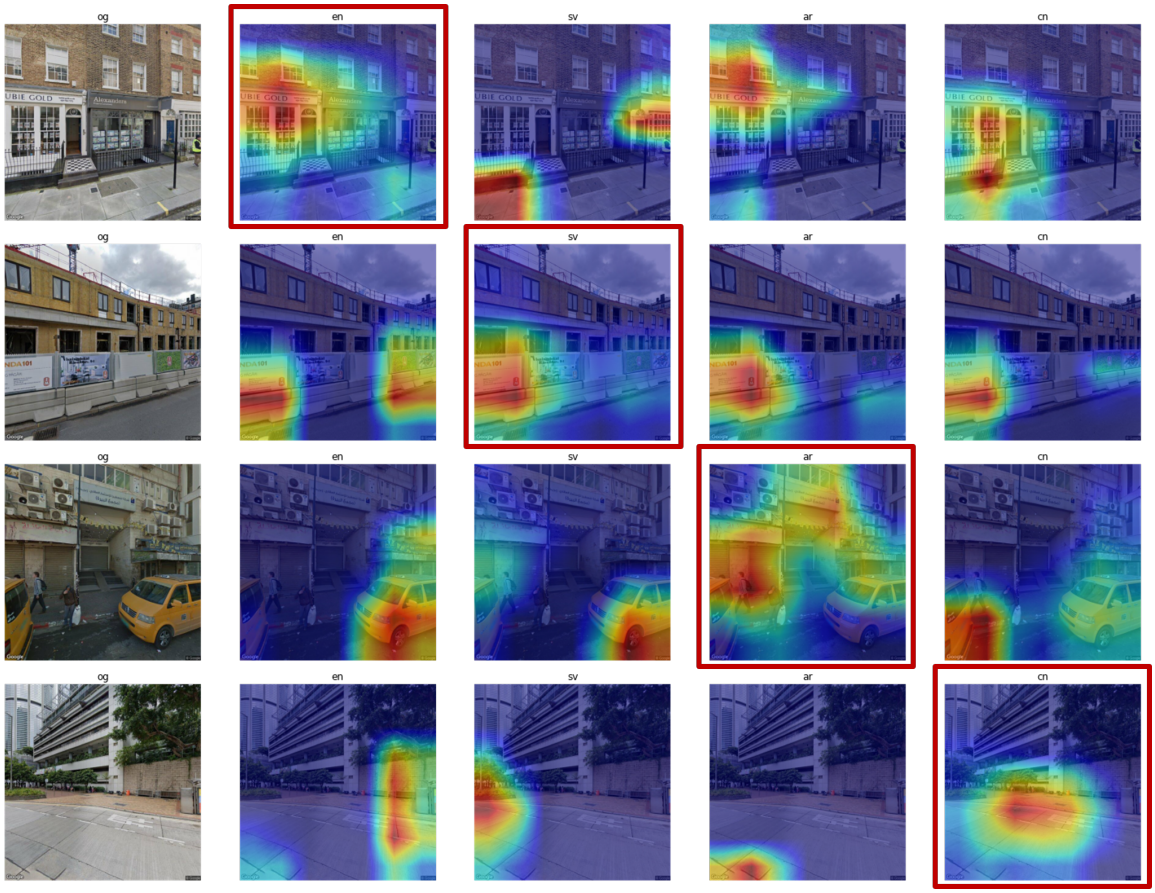

Figure 3-3: Grad-CAM performed for true positive examples. There is one example for English, Swedish, Arabic and Chinese (from top to bottom). Heatmaps for each classifier (English, Swedish, Arabic and Chinese, from left to right) are provided for each image. For each row, the heatmap highlighted in red is that of the classifier for which the image is a true positive example. In general, we see that the model is attentive to parts of the images for which there is text e.g. store signs or road markings.

3.5 Applications in Stockholm

3.5.1 Summary Statistics

We apply our classifier on GSV images in Stockholm, Sweden between 2009 and 2021. [8] To facilitate comparison with public socioeconomic data from Statistics Sweden [68] we aggregate our data at the DeSO (Demographic Statistical Areas) level. In Figure

[8] Details of data extraction are outlined in Appendix A.

3-6, we present choropleth maps of the linguistic concentration in Stockholm in 2020-2021.

We see that Swedish (unsurprisingly) has the strongest presence among the 4 languages across the city. English has a moderate presence, while Arabic and Chinese has minimal presence, and these three languages are more concentrated around the downtown area. As Hult observed in Sweden, English is not imposed from above but arises from socioeconomic interests [36]. It is therefore unsurprising to find a stronger presence of English in the downtown area where there are stronger commercial interests and higher tourist footfall.

3.5.2 Comparison with other Socioeconomic Characteristics

Statistics Sweden provides aggregated data of the citizenship of residents—residents are classified as "Swedish", "Europeans except Swedish", or "Others". To facilitate comparison with our measures of linguistic diversity, we construct a measure of population entropy in the same way we constructed our measure of linguistic entropy. In Figure 3-7, we present choropleth maps of linguistic entropy and population entropy in Stockholm in 2020-2021. We also include a choropleth map of median income to facilitate comparison.

Interestingly, and perhaps counterintuitively, the spatial distributions of the two measures are reversed, with areas that have higher linguistic entropy having lower population entropy. In fact, the correlation of the two measures across all years is -0.205, implying a weak negative correlation. Although this might not make sense at face value, this observation aligns well with our understanding of Swedish society. Crucially, we need to recognize that high linguistic mix need not arise from high population mix, particularly in a country whose heritage and culture is hallmarked by homogeneity. Rather, the linguistic diversity we see in Stockholm is not so much resultant of a diverse resident population along ethnolinguistic lines, but likely a feature of globalization. In fact, in a study of linguistic landscapes in Seoul, Hong

finds an increased prevalence of Chinese signs despite a relatively unchanged Chinese population and attributes this to "the recent popularity of Chinese food in the Korean society" [33], suggesting that linguistic diversity can be driven by preferences. One telltale sign for our case is the positive correlation between linguistic entropy and median income—0.261 which suggests that linguistic diversity may be well-associated with a diversity-seeking upper-middle class population that demands more culturally diverse goods and services.

On the other hand, despite the strong population entropy in the outskirts of the city, the lower linguistic entropy suggests that minorities are not necessarily comfortable expressing themselves in their mother tongues, in a country that prides itself on its cultural homogeneity. In fact, Daun notes that differences in cultural backgrounds are downplayed in accordance with the Swedish "emphasis on conflict avoidance" [17], and this may have contributed to the limited presence of minority languages, similar to the case of Polish in Norway.

3.6 Concluding Remarks

In this chapter, we introduce a different paradigm for detecting the presence of languages in streetscapes. Instead of using existing OCR tools, we propose the use of a pre-trained DenseNet-121 model to do binary classification for the presence of each language of interest. Our best model (which supports English, Swedish, Arabic and Chinese) achieves a test accuracy of 80.8%, surpassing the performance of existing OCR tools. We explore the use of both synthetic and real data in training our model and find that training solely on real data achieves the best performance, likely because synthetic data does not generalize well to real streetscapes and text in the wild. To check if the model is making sense, we employ Grad-CAM and find that the model is indeed focusing on visual features containing text. We apply our model to GSV images in Stockholm and find that the linguistic concentration of each of the 4 languages aligns with our intuition. We construct an entropy index to measure linguistic

diversity and compare it with population entropy. Although there is a weak negative correlation between the two, we argue that linguistic mix and population mix need not go hand-in-hand, particularly in a strongly homogeneous society like Sweden. Rather, the fact that linguistic mix is not commensurate with population mix points to the importance of measuring them separately.

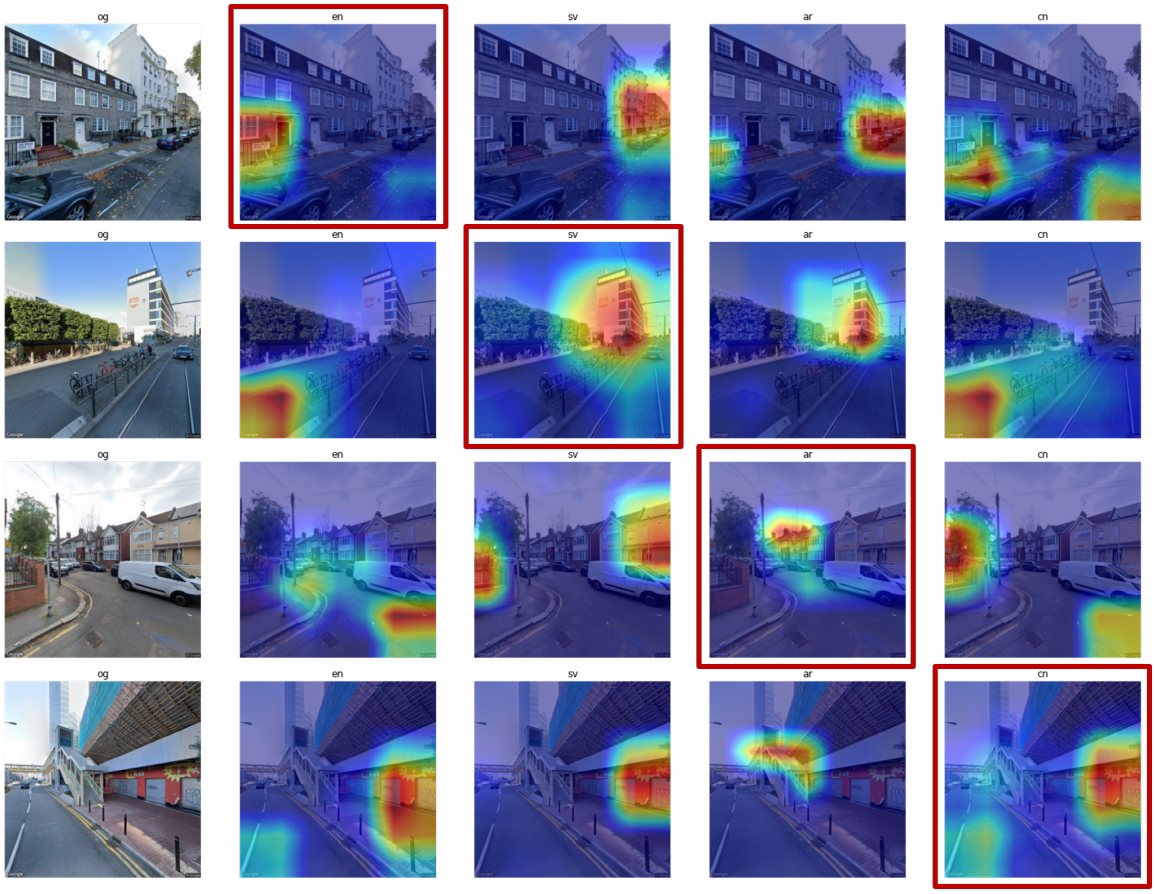

Figure 3-4: Grad-CAM performed for false negative examples. There is one example for English, Swedish, Arabic and Chinese (from top to bottom). Heatmaps for each classifier (English, Swedish, Arabic and Chinese, from left to right) are provided for each image. For each row, the heatmap highlighted in red is that of the classifier for which the image is a false negative example. In general, we see that the model fails to detect the presence of text despite being attentive to parts of the images for which there is text. We postulate that this is because the text is too small or indistinct from other visual features in a streetscape.

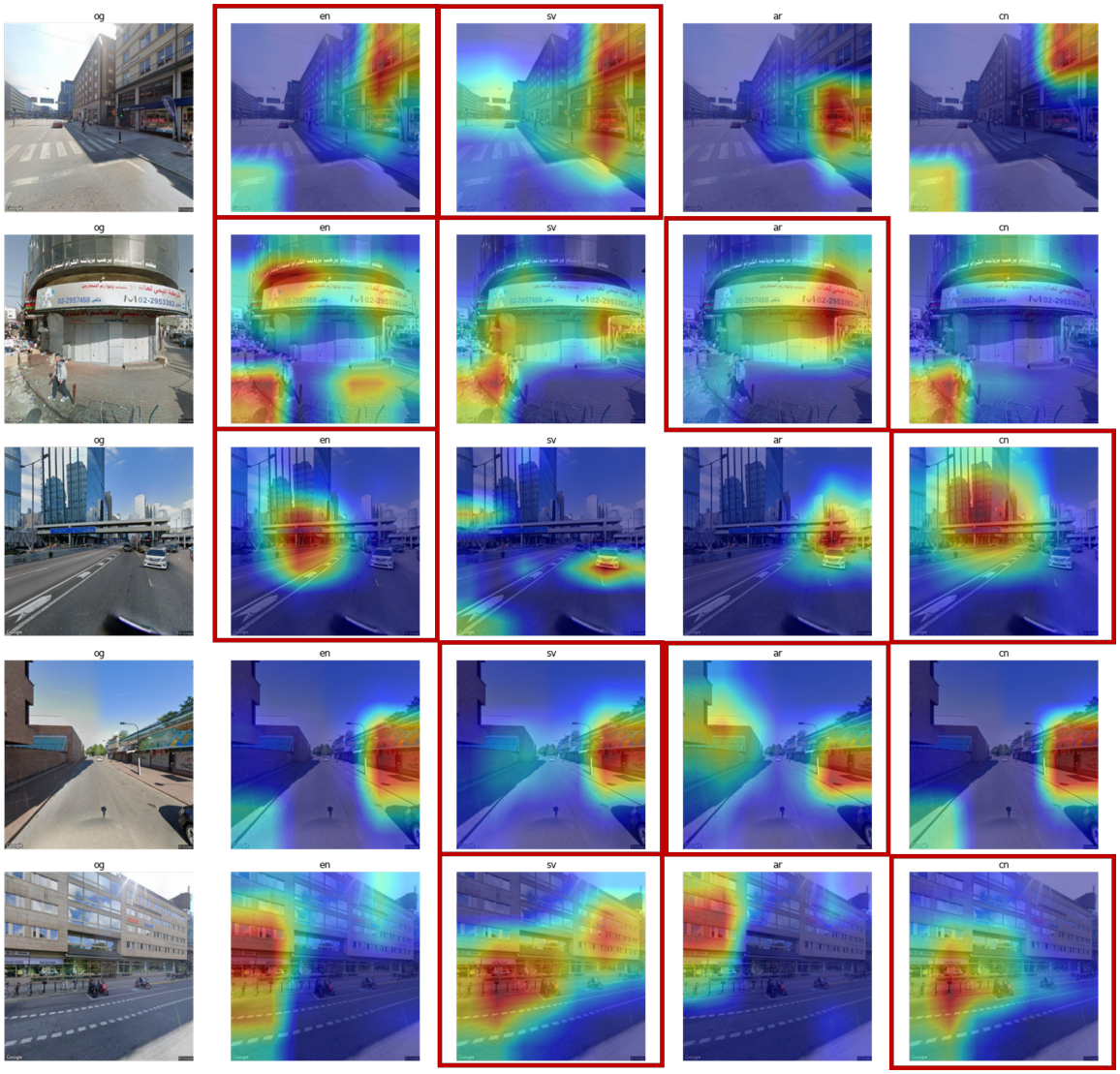

Figure 3-5: Grad-CAM performed for true positive multilingual examples. There is one example for English and Swedish, English and Arabic, English and Chinese, Swedish and Arabic, and Swedish and Chinese (from top to bottom). Heatmaps for each classifier (English, Swedish, Arabic and Chinese, from left to right) are provided for each image. For each row, the heatmaps highlighted in red are those of the classifiers for which the image is a true positive example. In general, we see that the model is attentive to parts of the images for which there is text e.g. store signs or road markings. In multilingual scenes, the pairs of classifiers tend to focus on similar parts of the image.

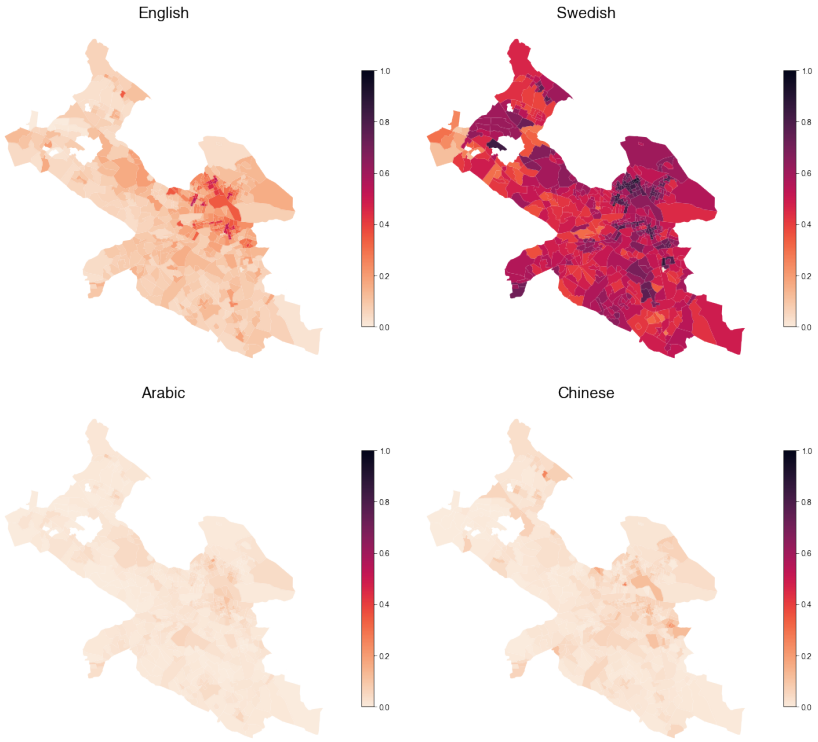

Figure 3-6: Linguistic concentration of English, Swedish, Arabic and Chinese in Stockholm in 2020-2021. The presence of Swedish is much higher than all other languages across the city. There is moderate presence of English in the downtown area while the presence of other foreign languages is limited.

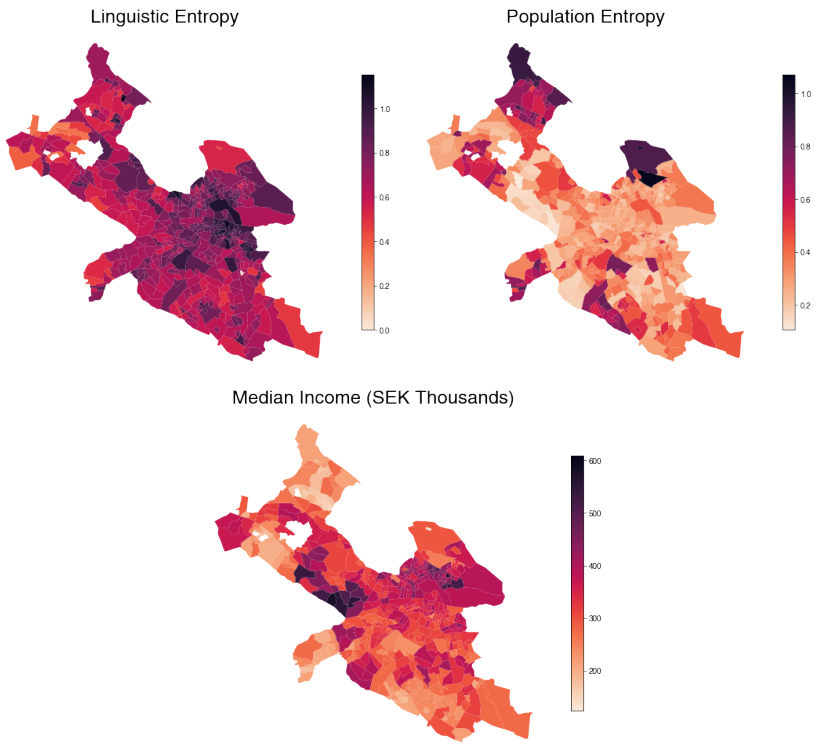

Figure 3-7: Linguistic entropy, population entropy and median income in Stockholm in 2020-2021

Chapter 4

City Change

Cities are molded by change—they are a tapestry of both public and private interventions in the urban space, and the rhythm of the city is captured through the changing interactions between people and place.

Urban interventions—new schools, housing, parks, bus networks, stadiums, public spaces—may have important intended and unintended implications for its immediate vicinity. Neighborhoods can become more vibrant due to the spillover effect of urban interventions—Hornbeck and Keniston noted how simultaneous reconstruction in areas afflicted by the Great Fire of Boston in 1872 set off a virtuous cycle of building upgrades even in areas unaffected by the Fire [34]. Jaime Lerner, former mayor of Curitiba, Brazil cited the construction of a provisional café that served as a new hub of activity as an example of filling in urban "voids", providing continuity in a space that was not devoid of people or housing, but of a space for interaction. Such examples are, as Lerner describes, *urban acupuncture*, "a spark that sets off a current that begins to spread" [45].

However, urban interventions can also foster gentrification processes or the degradation of the built environment, among other unintended consequences. It is not uncommon for existing residents to resist new developments within their neighbor-

hood for reasons such as short-term visual pollution, increased traffic noise, disruption to local traffic patterns, and long-term loss of a neighborhood's character. Therefore, it is key to understand the impacts of urban interventions and how these effects vary spatiotemporally.

In this chapter, we study how the construction of new housing projects affects the visual quality and linguistic diversity in Stockholm in the period 2009-2021. Stockholm is a growing city and its transformation over the years is characterized by a tireless process of construction and reconstruction—infill developments to increase the density of existing neighborhoods, new housing projects to expand the urban space and cater to the growth of a younger generation. Amidst expectations of continued growth, the city has set its sights on 140,000 new homes by 2030 and a corresponding expansion of its urban infrastructure [14]. Against this backdrop of growth and expected growth in Stockholm, it is important to understand what the consequences of new construction are. We find that a 1% increase in the number of housing projects within the 300m radius of a 100m $\times$ 100m grid can increase enclosure as measured by the proportion of pixels classified as walls or buildings by 0.0134% and linguistic entropy by 0.0375% in the period that construction is completed and the effects are persistent over time. These findings suggest that new housing construction can spur complementary developments in the surrounding area and also bring in people with a stronger desire for culturally diverse goods and services.

Our work contributes to a body of literature that estimates the externalities of urban interventions. Existing work in the literature has focused on the effects of urban interventions as measured by common socioeconomic indicators—neighboring land values [60], housing prices [18, 27], crime rate, income and racial diversity [18].

Studies focusing on housing or land values start by assuming a hedonic price model and see each housing unit as a basket of amenities and qualities. The general narrative is that changes in neighboring housing prices following an urban intervention arises from a confluence of demand and supply factors that can be driven by improve-

ments in building quality among other factors. In general, these studies estimate the relative impacts of supply and demand effects, but do not necessarily identify the mechanisms that underpin changes in supply and demand [3, 27, 69]. Although some studies measure changes in building quality more directly—e.g. Perkins et al. collate measures of self-reported home repairs and improvements and changes in the assessed conditions of yards and exteriors [58], such data is not easily available. This motivates the use of visual quality measures as a way for us to estimate the impact of new housing projects on the visual quality of neighboring urban environments.

In a country that has recently experienced large-scale immigration, the consequences of new housing projects on social integration, whether intended or unintended, is also of concern. Most studies look into the effects of new policies or projects on social segregation [26,50,63], which provides an objective measure of where people live, allowing us to then infer the level of integration. In Chapter 3, we argue that our measure of linguistic entropy is a measure of cultural capital that captures (1) the level of comfort that minorities have in showcasing their culture and (2) the demand among the general populace for multicultural goods and services. This measure goes beyond showing where people live but offers a perspective of the overall comfort with multiculturalism. Therefore, we also study the impact of new housing projects on the linguistic diversity in the surrounding environment.

Evaluating the effects of urban interventions on metrics constructed from street view imagery (SVI) follows from a body of literature that views cities as centers of aesthetic and recreational value and uses images to characterize cities [9, 24, 52]. While these studies demonstrate the usefulness of SVI in their ability to predict socioeconomic variables [9, 24][1] and vice versa [52], we go one step further by testing how metrics constructed with SVI change in response to urban interventions. We see this as an opportunity to evaluate how useful metrics constructed from SVI can be beyond the spatiotemporal characterization of urban areas.

[1]Technically, Carlino and Saiz only use *counts* of geotagged images, instead of the content of the images, like we do.

	Housing		
	Completed	Units	Rooms
2009	122	5039	19469
2010	103	3503	13382
2011	95	3184	12360
2012	89	4356	17141
2013	78	3604	13980
2014	84	3071	11474
2015	112	5206	19545
2016	86	4026	13114
2017	121	5162	18588
2018	102	5497	18512
2019	105	5413	18965
2020	81	4228	13888
2021	68	4519	13491

Table 4.1: Summary statistics of new construction of housing projects. Data is available for 2009-2021. We include the total number of projects, units and rooms completed in each year.

4.1 Data

4.1.1 Housing Projects

We obtain data of new housing projects from the City Planning Authority in Stockholm. The dataset includes all housing projects between 2009 and 2021. For each housing project, we are furnished with information of the number of units and rooms—information that can be used to account for varying treatment intensity. The summary statistics are presented in Table 4.1.

4.1.2 Dependent Variables

We obtain Google Street View (GSV) images of the streets of Stockholm from 2009 to 2021[2] and apply methods outlined in Chapter 2 and 3 to construct measures of

[2]Details of data extraction are outlined in Appendix A.

visual quality and linguistic diversity respectively. We aggregate data from every 2 years as a single temporal unit and data within each 100m $\times$ 100m grid as a single spatial unit to obtain a panel dataset. The summary statistics are outlined in Table 4.2.

	2009-10		2011-12		2013-14		2016-17		2018-19		2020-21	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
Imageability (People)	0.000293	0.00112	0.000375	0.00128	0.000486	0.00138	0.000565	0.00214	0.000479	0.00168	0.000313	0.000721
Imageability (SC, Top 5)	0.129	0.134	0.127	0.132	0.123	0.131	0.128	0.13	0.132	0.136	0.123	0.114
Imageability (SC, Prob)	0.0237	0.0184	0.0234	0.0182	0.023	0.018	0.0233	0.0177	0.0239	0.0184	0.0227	0.0157
Human Scale (Greenery)	0.135	0.0464	0.138	0.0451	0.138	0.0452	0.138	0.044	0.136	0.0461	0.139	0.0388
Human Scale (Furniture)	0.0194	0.012	0.019	0.0117	0.0191	0.0119	0.019	0.0114	0.0193	0.0119	0.0188	0.0101
Transparency	0.000995	0.00127	0.001	0.00122	0.000975	0.00119	0.000979	0.00126	0.00098	0.0012	0.00101	0.0012
Enclosure (Wall/Building)	0.0764	0.0779	0.0846	0.0828	0.104	0.0902	0.0937	0.0888	0.0867	0.0835	0.0822	0.0795
Enclosure (Sky)	0.114	0.053	0.108	0.0548	0.1	0.0579	0.113	0.058	0.125	0.0582	0.114	0.0599
Complexity (Objects)	9.88	1.06	9.9	1.06	9.88	1.05	9.86	1.01	9.89	1.06	9.89	0.922
Complexity (HH)	0.263	0.0724	0.265	0.0721	0.267	0.0734	0.265	0.0709	0.263	0.073	0.266	0.0641
Beautiful	5.94	0.451	5.96	0.439	5.96	0.445	5.96	0.424	5.94	0.451	5.97	0.384
Boring	5.58	0.211	5.58	0.208	5.59	0.208	5.59	0.2	5.58	0.208	5.59	0.184
Depressing	4.37	0.339	4.36	0.336	4.36	0.342	4.36	0.324	4.37	0.341	4.36	0.297
Lively	4.82	0.263	4.82	0.258	4.81	0.256	4.81	0.25	4.82	0.259	4.8	0.224
Safety	4.82	0.26	4.83	0.257	4.83	0.26	4.83	0.246	4.82	0.263	4.82	0.228
Wealthy	5.34	0.393	5.35	0.39	5.35	0.395	5.35	0.376	5.35	0.394	5.35	0.346
Linguistic Entropy	0.491	0.287	0.541	0.31	0.566	0.314	0.593	0.33	0.485	0.303	0.518	0.268

Table 4.2: Summary statistics of dependent variable grouped by time period

4.2 Methodology

4.2.1 Baseline

The standard approach for evaluating the effects of new construction in the vicinity is difference-in-difference (DID) [3, 18, 34, 46]. In the canonical DID model, we have 2 periods—*before* and *after* an intervention and 2 groups—a *treated* and a *control*. By assuming that the dependent variable in the treatment and control groups would have changed in parallel in the absence of treatment, we can estimate the average treatment effect on the treated by estimating the difference between the treatment group and the control group before and after an intervention. In the context of evaluating the effects of a new construction, the treatment group is construed as a ring around each intervention (*inner ring*) and the control group as a ring around the inner ring (*outer ring*).

In practice, we often have data across many time periods, which also creates variation in the treatment timing. This motivates a generalized DID model that is estimated using a dynamic two-way fixed effects (TWFE) regression specification.

$$Y_{igt} = \alpha_g + \phi_t + \sum_k \beta_k D_{i,g,t-k} + \varepsilon_{igt} \tag{4.1}$$

In our context, i is the grid ID, g the DeSO (Demographic Statistical Areas) ID, t the time period. α_g refers to spatial fixed effects, ϕ_t refers to time fixed effects and $D_{i,g,t-k}$ is a dummy variable that takes 1 if there is at least one project completed in period $t-k$ within a given distance of observation i. In this dynamic specification, we allow the causal estimand β_k to vary across time (relative to project completion). This allows us to tease out possible anticipation effects and evaluate how persistent treatment effects are.

Beyond temporal heterogeneity, recent work in urban economics and real estate literature also allow for spatial heterogeneity [7, 40, 57, 66]. Intuitively, we expect that

the treatment effect is stronger the closer a unit of observation is from an intervention. In these studies, the authors construct multiple treatment rings to estimate treatment effects at different distances from an intervention. This results in the following regression specification

$$Y_{igt} = \alpha_g + \phi_t + \sum_k \sum_r \beta_{k,r} D_{i,g,t-k,r} + \varepsilon_{igt} \tag{4.2}$$

where $D_{i,g,t-k,r}$ is a dummy variable that takes 1 if there is at least one project completed in period $t-k$ in the r-th ring of observation i. This yields a causal estimand $\beta_{k,r}$ that varies both spatially and temporally.

In the baseline, we use distance bins of 300m for the first 600m and 200m for the next 1.4km. Here, the identification assumption is that within a micro-neighborhood of 2km radius, the outcomes of interest would have changed in parallel in the absence of new construction. Intuitively, since the units of observations are within a short distance from one another, we expect that the only difference between these observations after controlling for fixed effects are their distances from new construction. We do not expect this hyperlocal variation to depend on the dependent variables nor other unobserved variables and exploit this variation to provide consistent estimates of treatment effects.

4.2.2 Variation in Treatment Intensity

Even with a generalized DID model that accounts for spatiotemporal heterogeneity in treatment effects (Equation 4.2), we do not account for varying treatment intensity arising from exposure to multiple interventions and project size. This is essential given how dense our dataset is.[3] There are two broad approaches in handling multiple treatments. Blanco and Neri allow for duplicate entries—i.e. observations exposed to multiple treatments appear multiple times in the dataset—and check for robustness by

[3] 96% of the observations are exposed to multiple treatments.

dropping different subsets of duplicates [7]. On the other hand, Pennington accounts for variation in treatment intensity by regressing the dependent variables of concern against the degree of exposure to new construction [57]. Therefore, we estimate the following equation in the baseline:

$$Y_{igt} = \alpha_g + \phi_t + \sum_k \sum_r \beta_{k,r} C_{i,g,t-k,r} + \varepsilon_{igt} \tag{4.3}$$

where instead of a dummy variable, we have $C_{i,g,t-k,r}$ on the right-hand-side to capture the number of projects completed in time $t-k$ in the r-th ring around grid i. As part of robustness checks, we define C as the number of units and the number of rooms too.

4.2.3 Variation by Income Group

Similar studies are often cognizant of heterogeneous effects in different parts of a city across socioeconomic lines. In particular, many studies have made the distinction between low-income and high-income areas. Asquith et al. noted that high-income areas may have better reputation, broader appeal, better amenities than low-income areas, with the interaction of underlying differences contributing to different empirical impacts and focused on low-income areas in their study of the effects of new large apartments [3]. Diamond and McQuade also made the distinction between high-income and low-income areas and found that properties financed by the Low Income Housing Tax Credit (LIHTC) have different qualitative impacts in high-income and low-income areas. Therefore, in our study, we also make a distinction between high-income and low-income areas [18]. We broadly define *high-income* areas as DeSO areas whose mean income exceeds the Stockholm mean income in the period of concern, and *low-income* areas as DeSO areas whose mean income falls below the Stockholm mean income in the period of concern. Although we expect differing impacts in high-income and low-income areas, we are agnostic about where these distinctions may fall given the confluence of factors that differ between high-income and low-income areas.

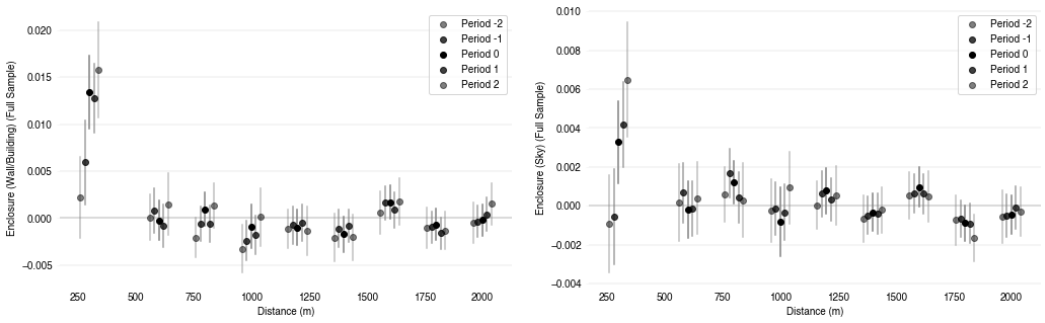

Figure 4-1: Event study plots of enclosure, using the full sample and the number of projects as the measure of treatment intensity

4.3 Results and Discussion

4.3.1 Visual Quality

In the baseline, we run the regression with a full sample of observations, using the number of housing projects in each ring as a measure of treatment intensity. Although the results are fairly noisy for the other dependent variables (Appendix C), we find that enclosure increases in the immediate vicinity (300m) of a new project and the effect is persistent over time. The graphs in Figure 4-1 show that a 1% increase in the number of housing projects completed within 300m of an area increases enclosure (wall/building) by 0.0134% and enclosure (sky) by 0.0033% in the time period the project is completed. The effect is persistent over time, and the estimated coefficients rises to 0.0157 for enclosure (wall/building) and to 0.0065 for enclosure (sky) 2 periods after the project is completed.

Intuitively, the completion of new housing projects may spur complementary developments such as increased commercial space or building improvements in the surrounding area, thereby accounting for increased enclosure. This aligns with Perkins et al., who found modest effects on a series of building quality measures within one block of new housing [58].

When we run the regressions for samples segregated by income levels, we find that

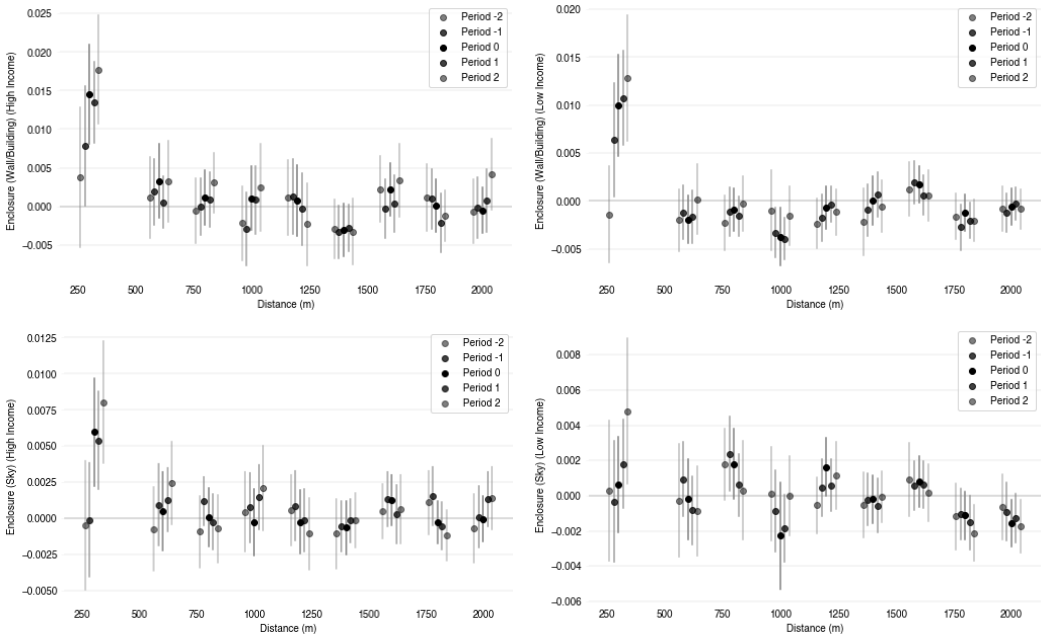

Figure 4-2: Event study plots of enclosure, using samples segregated by income and the number of projects as the measure of treatment intensity

the effects on enclosure are seen in both high-income and low-income areas, with the estimated effects being generally larger in high-income areas (Figure 4-2).

However, when we focus specifically on low-income areas, we see that the immediate vicinity sees an increase in imageability (people), becomes less depressing but more boring (Figure 4-3). Although it is unclear why such effects are more pronounced in low-income areas, it makes sense that a new housing project should increase footfall in the area and the completion of construction and removing of scaffolding should make a place look less depressing. Even though we expect an area to look less boring with urban development in general, the increase in enclosure may account for urban areas feeling more stifling and "boring". In a study to uncover physical features that can explain perceptual features, Zhang et al. argue that walls may lead to blocked views, decreased sunshine, and the build-up of pollution, resulting in scenes with a large proportion of walls to generate feelings of boredom [74].

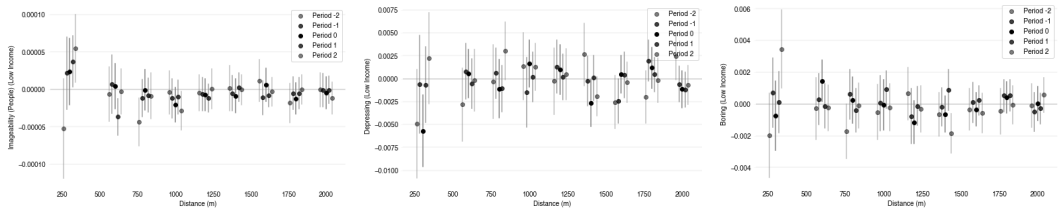

Figure 4-3: Event study plots of imageability (people), depressing and boring, using the low-income sample and the number of projects as the measure of treatment intensity

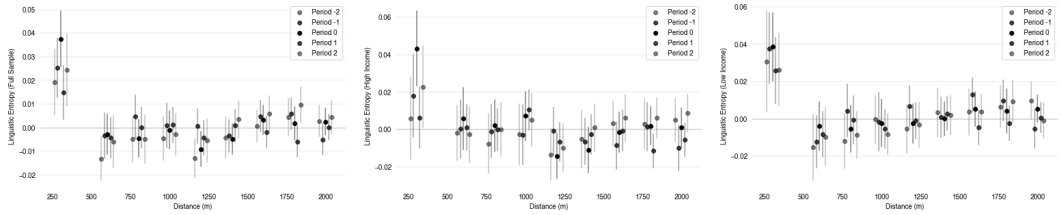

Figure 4-4: Event study plots of linguistic entropy, using the number of projects as the measure of treatment intensity

4.3.2 Linguistic Entropy

In the baseline, we find that there is a strong increase in linguistic entropy in the immediate vicinity of the construction. Based on Figure 4-4, we see that entropy rises by 0.0194% for every 1% increase in the number of housing projects 2 periods before project completion. The effect peaks in period 0 at 0.0375% but remains positive and statistically significant at 0.0244% in period 2, demonstrating temporal persistence. These effects are seen in both high-income and low-income areas (Figure 4-4).

In Chapter 3, we argue that linguistic entropy is not so much a measure of population mix. Rather, it is a measure of interest *per se* that tells us how diverse the linguistic landscape is. As we show in Section 3.5.2, linguistic entropy has a stronger (and positive) correlation with median income than with population mix. Therefore, we do not see linguistic entropy as a result of a more diverse residential population. Rather, we consider it an outcome of a stronger demand for culturally diverse goods and services from people who reside in and frequent an area. Therefore, we interpret

the increase in linguistic entropy in the vicinity of a new housing project following its completion as the outcome of a rise in a population demanding more culturally diverse products.

Crucially, several academics have characterized housing projects in Sweden today as luxury goods that serve as an expression of one's lifestyle [11,26,32]. Therefore, it is likely that these new housing projects attract middle-income, high-income-types with a stronger cultural consciousness and an appetite for diversity into the neighborhood, generating the demand that results in a stronger linguistic mix in the neighborhood. We also find that the effects in low-income areas are generally stronger than those in high-income areas.

4.4 Robustness Checks

4.4.1 Varying Definitions of Spatial Bins

As a robustness check, we use larger spatial bins in our regressions. Using larger spatial bins means that there is more data for each bin and this would allow us to estimate the coefficients more precisely and check if zero-effects at further distances are indeed zero. However, this comes at the expense of the spatial granularity of treatment effects.

We find that the effects of housing construction on enclosure (Figure 4-5) and linguistic entropy (Figure 4-6) remain strong in the immediate vicinity. However, the estimated coefficients are now smaller—the estimates in period 0 on enclosure (wall/building), enclosure (sky) and entropy are now 0.0072, 0.0017, 0.0179 respectively. This makes sense if the effects exhibit distance decay—since the smallest ring now includes projects further away from the unit of observation than before, the average treatment effect is diluted.

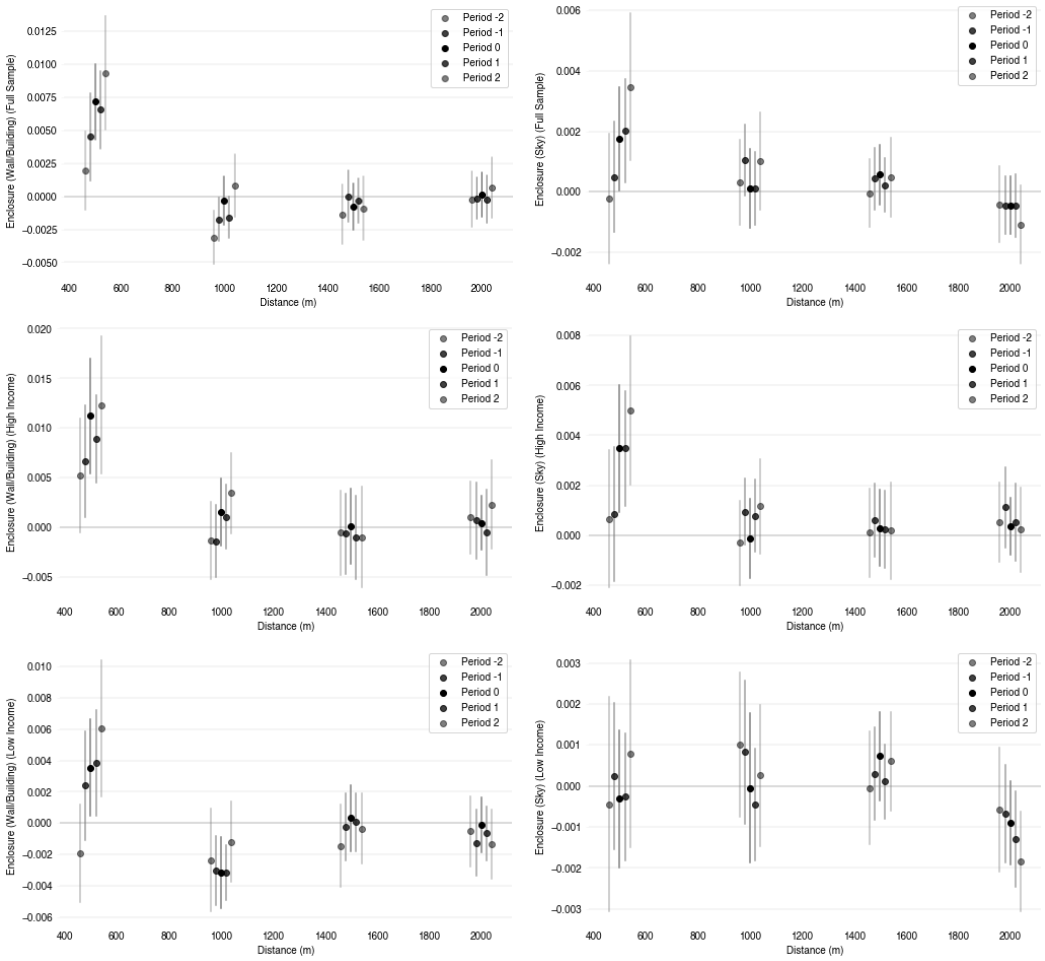

Figure 4-5: Event study plots of enclosure, using the number of projects as the measure of treatment intensity, with 500m spatial bins

4.4.2 Varying Definitions of Treatment Intensity

Since we are furnished with data of the scale of the different housing projects in our study, we repeat the study for different definitions of treatment intensity—using the number of units and number of rooms instead of the number of projects.

We find similar results regardless of definition (Appendix C). Intuitively, there is a strong positive correlation between the number of projects and the number of units or rooms in each ring, thereby making it likely for us to see the same qualitative results. In general, however, we find that the estimated coefficients are now smaller. This is intuitive too, since we expect the effect of a 1% increase in the number of rooms or

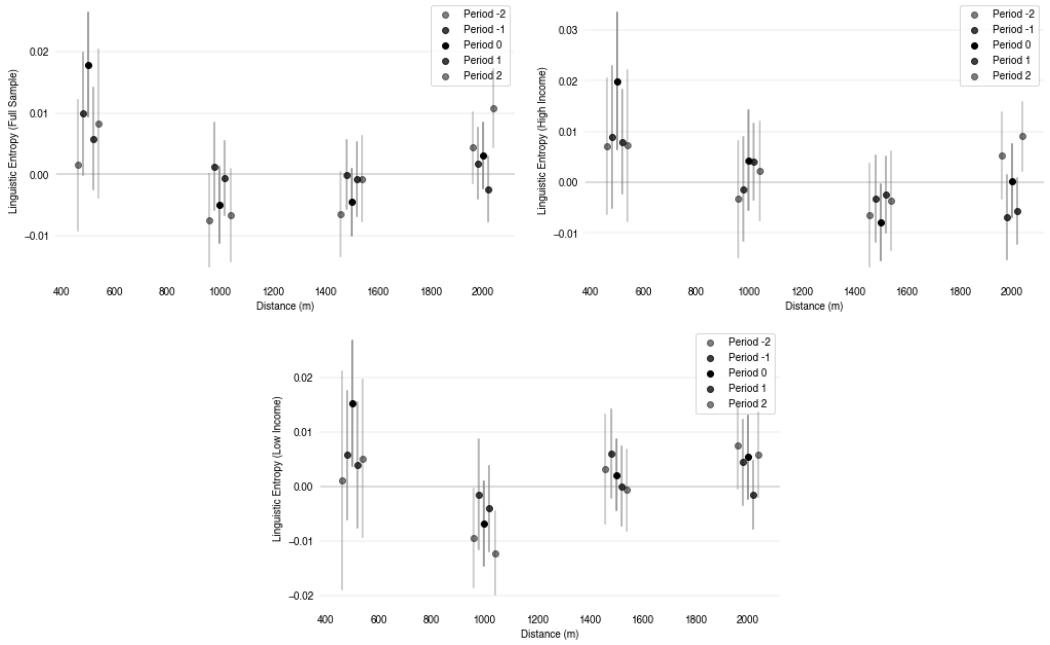

Figure 4-6: Event study plots of linguistic entropy, using the number of projects as the measure of treatment intensity, with 500m spatial bins

units to be smaller than the effect of a 1% increase in the number of projects.

4.5 Concluding Remarks

In this chapter, we apply novel metrics constructed in Chapters 2 and 3 to conduct a study of the effects of new housing projects, and how these effects vary spatiotemporally. A number of the measures constructed yield noisy estimates in a DID setting, which points to the limited effectiveness of measures constructed from SVI beyond spatiotemporal characterization. That said, we derive some intuitive results, showing that enclosure and linguistic entropy both increase in the immediate vicinity of new construction, with effects that are persistent over at least two periods (of two years) after construction. We interpret the increase in enclosure as an increase in complementary development in the surrounding area of new housing construction, and the increase in linguistic entropy as an increase in a resident population that is more culturally conscious and demands more cultural diversity. These results are robust

to different spatial bins and different measures of treatment intensity.

Chapter 5

Conclusion

In Chapter 1, we highlighted how the growth of street view imagery (SVI) has both created a wealth of data in the urban sphere and contributed to the rise of the deep learning paradigm. In this thesis, we leveraged on these developments to construct novel metrics—visual quality and linguistic diversity to (1) characterize urban streetscapes spatiotemporally and (2) apply them in studying the effects of constructing new housing projects. We find that the metrics we have constructed are generally able to offer insightful spatiotemporal characterizations of our case study of Stockholm when used in conjunction with existing metrics, and some measures are also able to capture intuitive relationships when used in a difference-in-difference (DID) framework. There are limitations to these measures, because (1) they are ultimately quantifying abstract concepts and (2) they can only be as accurate as the underlying accuracy of the predictive models used to construct them. We summary the key findings of each chapter below:

5.1 Key Findings

5.1.1 Quantifying Visual Quality

In Chapter 2, we showed how we can leverage on modern computer vision tools to construct measures of visual quality. We used a scene classifier trained on the Places Database and a semantic segmentation model trained on the ADE20K dataset to construct measures of physical features that characterize the visual quality of an urban environment. We then trained a multiclassification model on the Place Pulse 2.0 dataset to construct perception scores. Applying these models to Stockholm, we found that our measures of visual quality generally demonstrate intuitive relationships with population density and median income, and with one another. However, we are also aware of the limitations of applying imperfect models to construct these measures. In particular, we found that measures constructed by the scene classifier are less useful than those constructed by the semantic segmentation model, which follows from the fact that it is more difficult to classify abstract notions (scenes) than clearly defined objects.

5.1.2 Quantifying Linguistic Diversity

In Chapter 3, we showed how modern computer vision tools can be used to detect the presence of languages in streetscapes, allowing us to then construct a second-order metric of linguistic diversity. We constructed a large dataset of SVI scraped from Google Street View (GSV), labelled with the presence of English, Swedish, Arabic and Chinese. We then trained a DenseNet-121 model on this dataset and the test accuracy of our best-performing model surpasses the performance of existing OCR tools on the same dataset. We applied the model to Stockholm and created an entropy measure based on the linguistic concentration in each 100m $\times$ 100m unit, finding that linguistic mix is negatively correlated with population mix and positively correlated with median income. This finding suggests that linguistic mix in Stockholm is not

driven by population mix, but likely by a demand for culturally diverse goods and services. Furthermore, given that linguistic mix is not commensurate with population mix, we argued that minorities may not be comfortable in showcasing their culture, even in areas with higher population mix. These findings point to how insightful linguistic diversity is as a measure, and the importance of measuring it separately.

5.1.3 Effects of Urban Interventions on Visual Quality and Linguistic Diversity

Given the insights offered by the measures of visual quality and linguistic diversity in the spatiotemporal characterization of Stockholm, we study the effects of new housing construction in Stockholm between 2009 and 2021 as measured by these metrics in Chapter 4. Although most measures of visual quality demonstrated intuitive spatiotemporal relationships, we found that they are less useful in a DID setting, with only the measures of enclosure yielding intuitive results. We find that both enclosure and linguistic diversity exhibit persistent increases in the immediate vicinity of new construction, capturing modest spillover effects in the neighborhood of new housing projects. In the larger context of quantifying urban streetscapes, these encouraging findings also highlight the potential of metrics constructed from SVI in helping us uncover more nuanced and robust causal relationships in the urban sphere. As more data and more powerful prediction models emerge, we expect metrics constructed from SVI to play an even more important role in understanding cities.

Appendix A

Google Street View in Stockholm

A.1 Querying Process

We begin by generating sampling points along the road network in Stockholm at 50-meter intervals using OpenStreetMap. We then make API requests from Google Street View (GSV) using these sampled coordinates and the following parameters—90° field-of-vision, 0° pitch, 50m radius. For each set of coordinates, we obtain images at compass headings of 0°, 90°, 180° and 270°, thereby capturing the full panorama at each point.

A.2 Summary Statistics

We obtain images from the period 2009-2021 and this amounts to 1026960 unique images, corresponding to 256740 unique panoramas. Since Google only conducts a large-scale update of images approximately every two years, we group the data in sets of two years (skipping 2015 as there is no data available for 2015). Furthermore, since the panorama for a given area has a unique ID that varies temporally, we aggregate our data at 100m × 100m grids in order to obtain a panel dataset. We present the

data availability for GSV images in Figure A-1.

Figure A-1: GSV data availability in Stockholm. Grids (100m $\times$ 100m) with images present for the particular year range are highlighted in blue.

Appendix B

Categories in Machine Learning Models

B.1 ADE20K Categories

The ADE20K scene parsing dataset used to train the semantic segmentation model comprises the following 150 classes (Table B.1):

1	wall	2	building	3	sky	4	floor
5	tree	6	ceiling	7	road	8	bed
9	windowpane	10	grass	11	cabinet	12	sidewalk
13	person	14	earth	15	door	16	table
17	mountain	18	plant	19	curtain	20	chair
21	car	22	water	23	painting	24	sofa
25	shelf	26	house	27	sea	28	mirror
29	rug	30	field	31	armchair	32	seat
33	fence	34	desk	35	rock	36	wardrobe
37	lamp	38	bathtub	39	railing	40	cushion
41	base	42	box	43	column	44	signboard
45	chest	46	counter	47	sand	48	sink
49	skyscraper	50	fireplace	51	refrigerator	52	grandstand

53	path	54	stairs	55	runway	56	case
57	pool	58	pillow	59	screen	60	stairway
61	river	62	bridge	63	bookcase	64	blind
65	coffee	66	toilet	67	flower	68	book
69	hill	70	bench	71	countertop	72	stove
73	palm	74	kitchen	75	computer	76	swivel
77	boat	78	bar	79	arcade	80	hovel
81	bus	82	towel	83	light	84	truck
85	tower	86	chandelier	87	awning	88	streetlight
89	booth	90	television	91	airplane	92	dirt
93	apparel	94	pole	95	land	96	bannister
97	escalator	98	ottoman	99	bottle	100	buffet
101	poster	102	stage	103	van	104	ship
105	fountain	106	conveyer	107	canopy	108	washer
109	plaything	110	swimming	111	stool	112	barrel
113	basket	114	waterfall	115	tent	116	bag
117	minibike	118	cradle	119	oven	120	ball
121	food	122	step	123	tank	124	trade
125	microwave	126	pot	127	animal	128	bicycle
129	lake	130	dishwasher	131	screen	132	blanket
133	sculpture	134	hood	135	sconce	136	vase
137	traffic	138	tray	139	ashcan	140	fan
141	pier	142	screen	143	plate	144	monitor
145	bulletin	146	shower	147	radiator	148	glass
149	clock	150	flag				

Table B.1: Classes covered in the ADE20K dataset

B.2 Places Categories

The Places Database used to train the scene classifier comprises the following 365 categories (Table B.2):

1	airfield	2	airplane_cabin	3	airport_terminal

4	alcove	5	alley	6	amphitheater
7	amusement_arcade	8	amusement_park	9	apartment_building (outdoor)
10	aquarium	11	aqueduct	12	arcade
13	arch	14	archaelogical_excavation	15	archive
16	arena/hockey	17	arena/performance	18	arena/rodeo
19	army_base	20	art_gallery	21	art_school
22	art_studio	23	artists_loft	24	assembly_line
25	athletic_field (outdoor)	26	atrium/public	27	attic
28	auditorium	29	auto_factory	30	auto_showroom
31	badlands	32	bakery/shop	33	balcony/exterior
34	balcony/interior	35	ball_pit	36	ballroom
37	bamboo_forest	38	bank_vault	39	banquet_hall
40	bar	41	barn	42	barndoor
43	baseball_field	44	basement	45	basketball_court (indoor)
46	bathroom	47	bazaar/indoor	48	bazaar/outdoor
49	beach	50	beach_house	51	beauty_salon
52	bedchamber	53	bedroom	54	beer_garden
55	beer_hall	56	berth	57	biology_laboratory
58	boardwalk	59	boat_deck	60	boathouse
61	bookstore	62	booth/indoor	63	botanical_garden
64	bow_window (indoor)	65	bowling_alley	66	boxing_ring
67	bridge	68	building_facade	69	bullring
70	burial_chamber	71	bus_interior	72	bus_station/indoor
73	butchers_shop	74	butte	75	cabin/outdoor
76	cafeteria	77	campsite	78	campus
79	canal/natural	80	canal/urban	81	candy_store
82	canyon	83	car_interior	84	carrousel
85	castle	86	catacomb	87	cemetery
88	chalet	89	chemistry_lab	90	childs_room
91	church/indoor	92	church/outdoor	93	classroom
94	clean_room	95	cliff	96	closet
97	clothing_store	98	coast	99	cockpit

100	coffee_shop	101	computer_room	102	conference_center
103	conference_room	104	construction_site	105	corn_field
106	corral	107	corridor	108	cottage
109	courthouse	110	courtyard	111	creek
112	crevasse	113	crosswalk	114	dam
115	delicatessen	116	department_store	117	desert/sand
118	desert/vegetation	119	desert_road	120	diner/outdoor
121	dining_hall	122	dining_room	123	discotheque
124	doorway/outdoor	125	dorm_room	126	downtown
127	dressing_room	128	driveway	129	drugstore
130	elevator/door	131	elevator_lobby	132	elevator_shaft
133	embassy	134	engine_room	135	entrance_hall
136	escalator/indoor	137	excavation	138	fabric_store
139	farm	140	fastfood_restaurant	141	field/cultivated
142	field/wild	143	field_road	144	fire_escape
145	fire_station	146	fishpond	147	flea_market/indoor
148	florist_shop/indoor	149	food_court	150	football_field
151	forest/broadleaf	152	forest_path	153	forest_road
154	formal_garden	155	fountain	156	galley
157	garage/indoor	158	garage/outdoor	159	gas_station
160	gazebo/exterior	161	general_store/indoor	162	general_store/outdoor
163	gift_shop	164	glacier	165	golf_course
166	greenhouse/indoor	167	greenhouse/outdoor	168	grotto
169	gymnasium/indoor	170	hangar/indoor	171	hangar/outdoor
172	harbor	173	hardware_store	174	hayfield
175	heliport	176	highway	177	home_office
178	home_theater	179	hospital	180	hospital_room
181	hot_spring	182	hotel/outdoor	183	hotel_room
184	house	185	hunting_lodge/outdoor	186	ice_cream_parlor
187	ice_floe	188	ice_shelf	189	ice_skating_rink (indoor)
190	ice_skating_rink (outdoor)	191	iceberg	192	igloo
193	industrial_area	194	inn/outdoor	195	islet
196	jacuzzi/indoor	197	jail_cell	198	japanese_garden
199	jewelry_shop	200	junkyard	201	kasbah

202	kennel/outdoor	203	kindergarden_classroom	204	kitchen
205	lagoon	206	lake/natural	207	landfill
208	landing_deck	209	laundromat	210	lawn
211	lecture_room	212	legislative_chamber	213	library/indoor
214	library/outdoor	215	lighthouse	216	living_room
217	loading_dock	218	lobby	219	lock_chamber
220	locker_room	221	mansion	222	manufactured_home
223	market/indoor	224	market/outdoor	225	marsh
226	martial_arts_gym	227	mausoleum	228	medina
229	mezzanine	230	moat/water	231	mosque/outdoor
232	motel	233	mountain	234	mountain_path
235	mountain_snowy	236	movie_theater/indoor	237	museum/indoor
238	museum/outdoor	239	music_studio	240	natural_history_museum
241	nursery	242	nursing_home	243	oast_house
244	ocean	245	office	246	office_building
247	office_cubicles	248	oilrig	249	operating_room
250	orchard	251	orchestra_pit	252	pagoda
253	palace	254	pantry	255	park
256	parking_garage (indoor)	257	parking_garage (outdoor)	258	parking_lot
259	pasture	260	patio	261	pavilion
262	pet_shop	263	pharmacy	264	phone_booth
265	physics_laboratory	266	picnic_area	267	pier
268	pizzeria	269	playground	270	playroom
271	plaza	272	pond	273	porch
274	promenade	275	pub/indoor	276	racecourse
277	raceway	278	raft	279	railroad_track
280	rainforest	281	reception	282	recreation_room
283	repair_shop	284	residential_neighborhood	285	restaurant
286	restaurant_kitchen	287	restaurant_patio	288	rice_paddy
289	river	290	rock_arch	291	roof_garden
292	rope_bridge	293	ruin	294	runway
295	sandbox	296	sauna	297	schoolhouse
298	science_museum	299	server_room	300	shed
301	shoe_shop	302	shopfront	303	shopping_mall/indoor
304	shower	305	ski_resort	306	ski_slope

307	sky	308	skyscraper	309	slum
310	snowfield	311	soccer_field	312	stable
313	stadium/baseball	314	stadium/football	315	stadium/soccer
316	stage/indoor	317	stage/outdoor	318	staircase
319	storage_room	320	street	321	subway_station/platform
322	supermarket	323	sushi_bar	324	swamp
325	swimming_hole	326	swimming_pool/indoor	327	swimming_pool/outdoor
328	synagogue/outdoor	329	television_room	330	television_studio
331	temple/asia	332	throne_room	333	ticket_booth
334	topiary_garden	335	tower	336	toyshop
337	train_interior	338	train_station/platform	339	tree_farm
340	tree_house	341	trench	342	tundra
343	underwater (ocean_deep)	344	utility_room	345	valley
346	vegetable_garden	347	veterinarians_office	348	viaduct
349	village	350	vineyard	351	volcano
352	volleyball_court (outdoor)	353	waiting_room	354	water_park
355	water_tower	356	waterfall	357	watering_hole
358	wave	359	wet_bar	360	wheat_field
361	wind_farm	362	windmill	363	yard
364	youth_hostel	365	zen_garden		

Table B.2: Scene categories covered in the Places365 dataset

Appendix C

Omitted Results

C.1 Baseline Regressions

In the baseline, we estimate:

$$Y_{igt} = \alpha_g + \phi_t + \sum_k \sum_r \beta_{k,r} C_{i,g,t-k,r} + \varepsilon_{igt} \tag{C.1}$$

where instead of a dummy variable, we have $C_{i,g,t-k,r}$ on the right-hand-side to capture the number of projects completed in time $t-k$ in the r-th ring around grid i. In Figures C-1, C-2 and C-3, we present the baseline event study plots of all dependent variables other than those already presented in Chapter 4, using the full sample, high-income sample, low-income sample respectively.

C.1.1 Full Sample

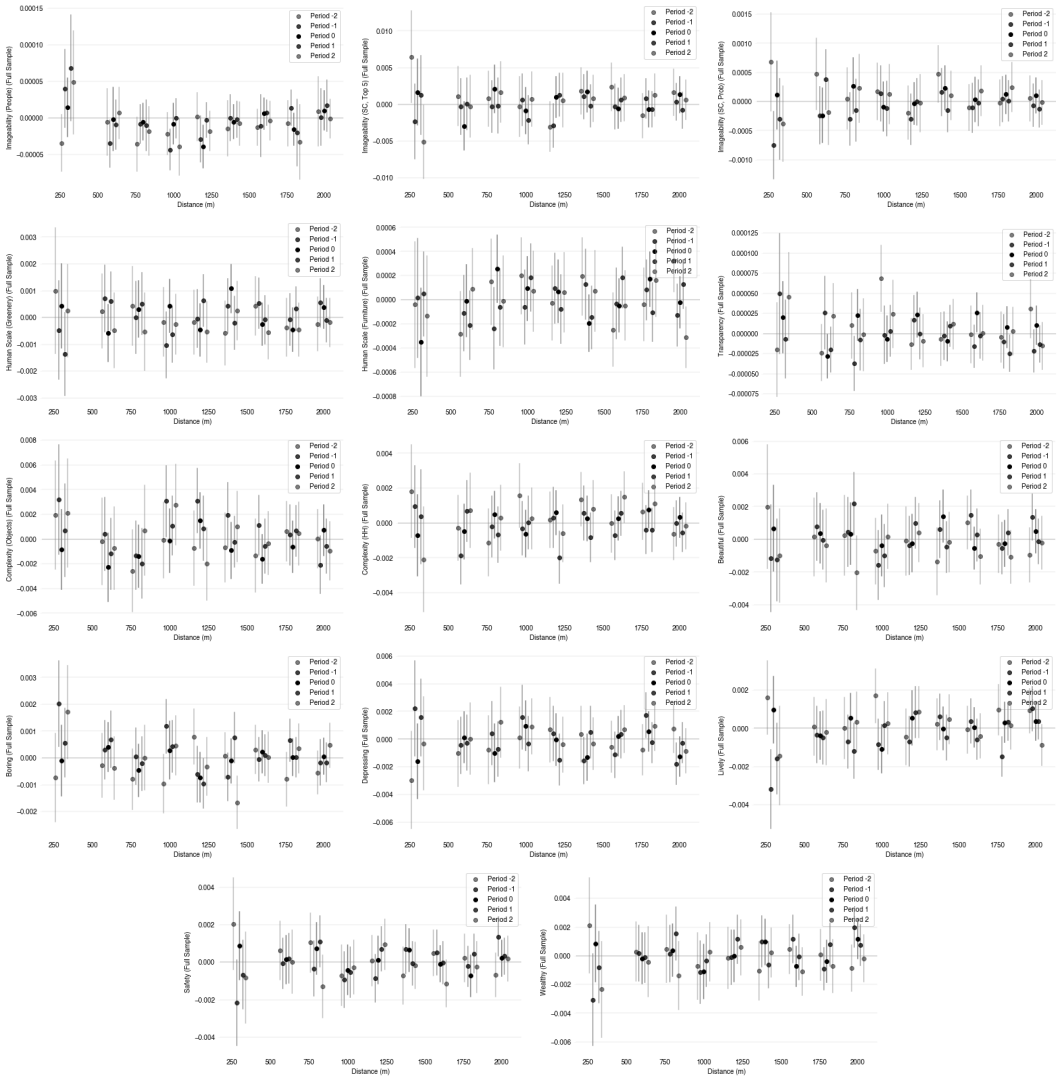

Figure C-1: Event study plots of all dependent variables other than enclosure and entropy, using the full sample and the number of projects as the measure of treatment intensity

C.1.2 High-Income Areas

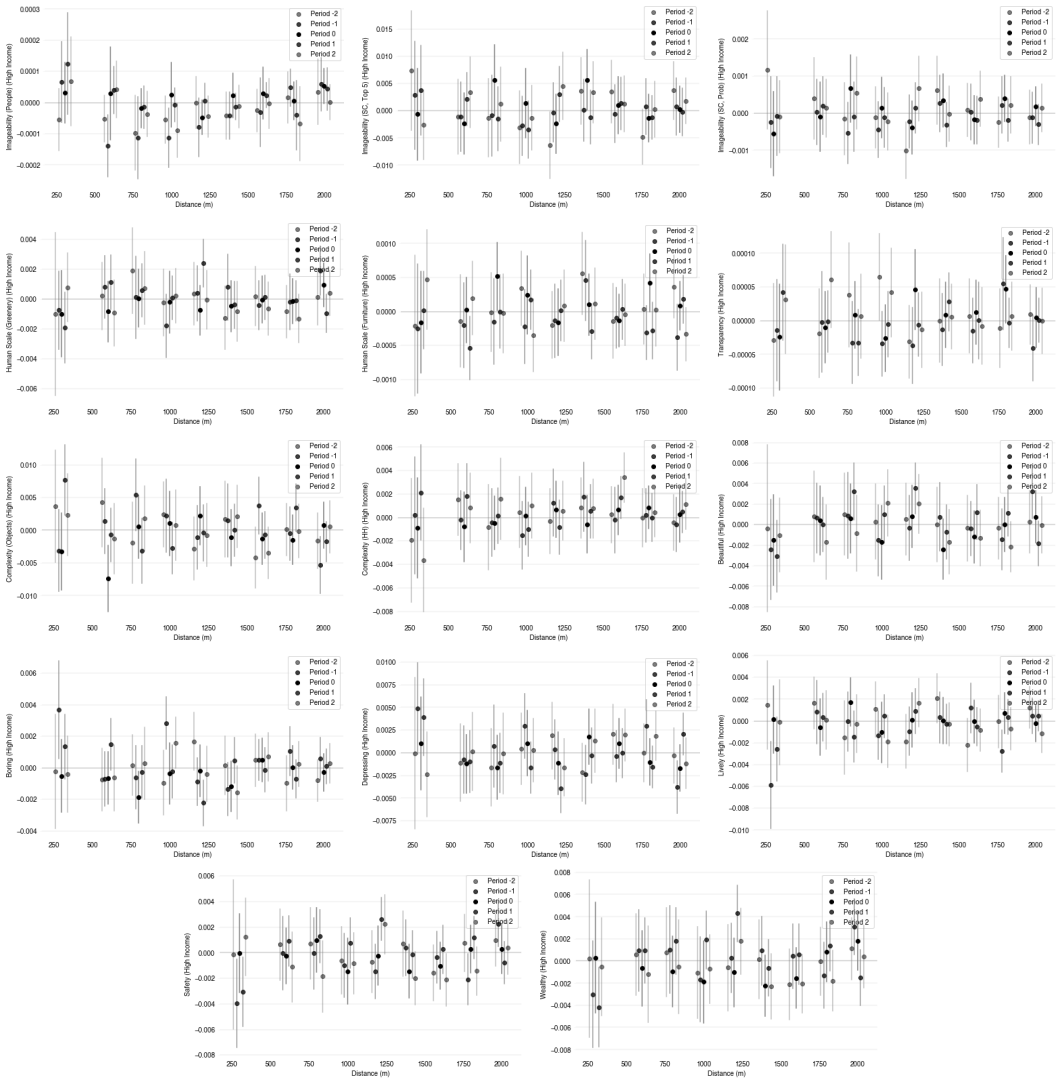

Figure C-2: Event study plots of all dependent variables other than enclosure and entropy, using the high-income sample and the number of projects as the measure of treatment intensity

C.1.3 Low-Income Areas

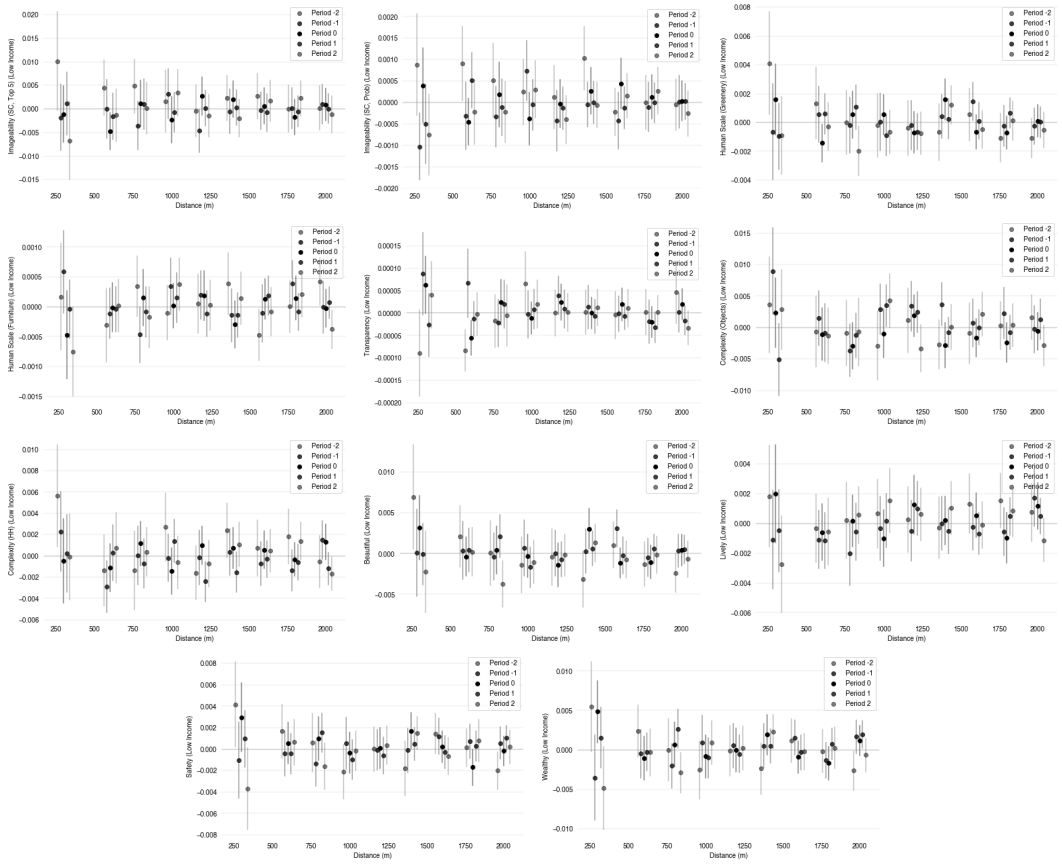

Figure C-3: Event study plots of all dependent variables other than enclosure, entropy, imageability (people), depressing and boring, using the low-income sample and the number of projects as the measure of treatment intensity

C.2 Robustness Checks (Treatment Intensity Measures)

In the figures below, we present event study plots corresponding to positive results in Section 4.3 using different measures of treatment intensity.

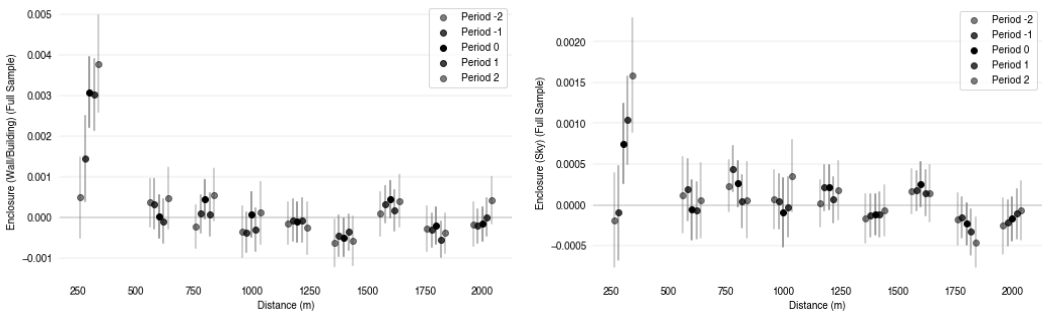

Figure C-4: Effects on enclosure in the full sample using the number of units as the treatment intensity measure

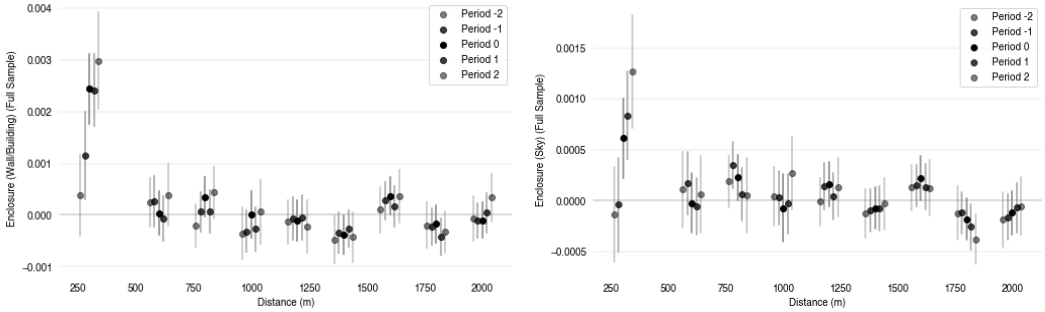

Figure C-5: Effects on enclosure in the full sample using the number of rooms as the treatment intensity measure

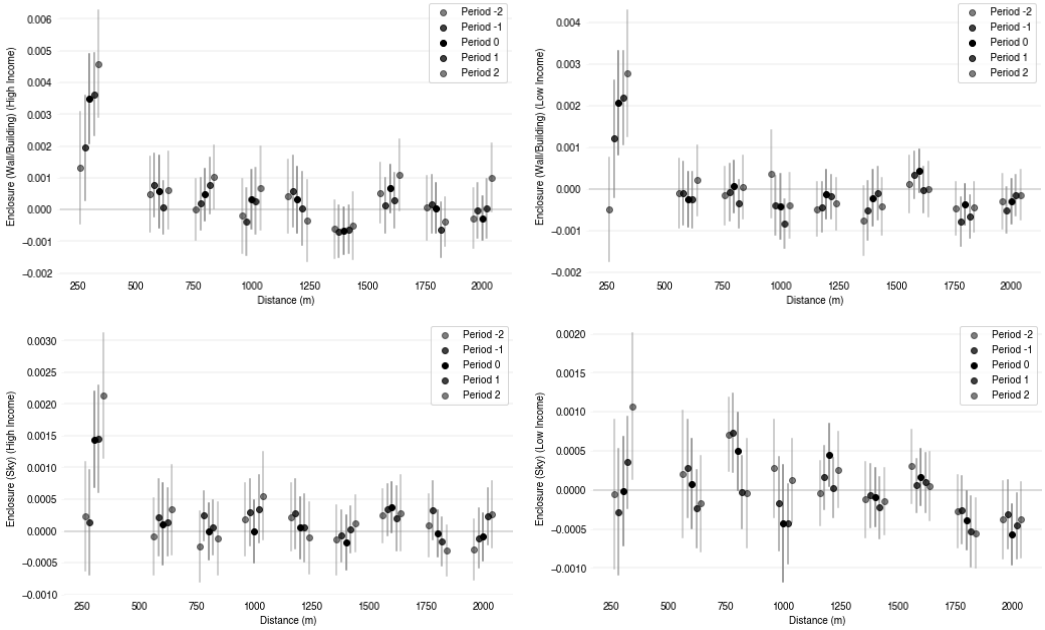

Figure C-6: Effects on enclosure in high- and low-income areas using the number of units as the treatment intensity measure

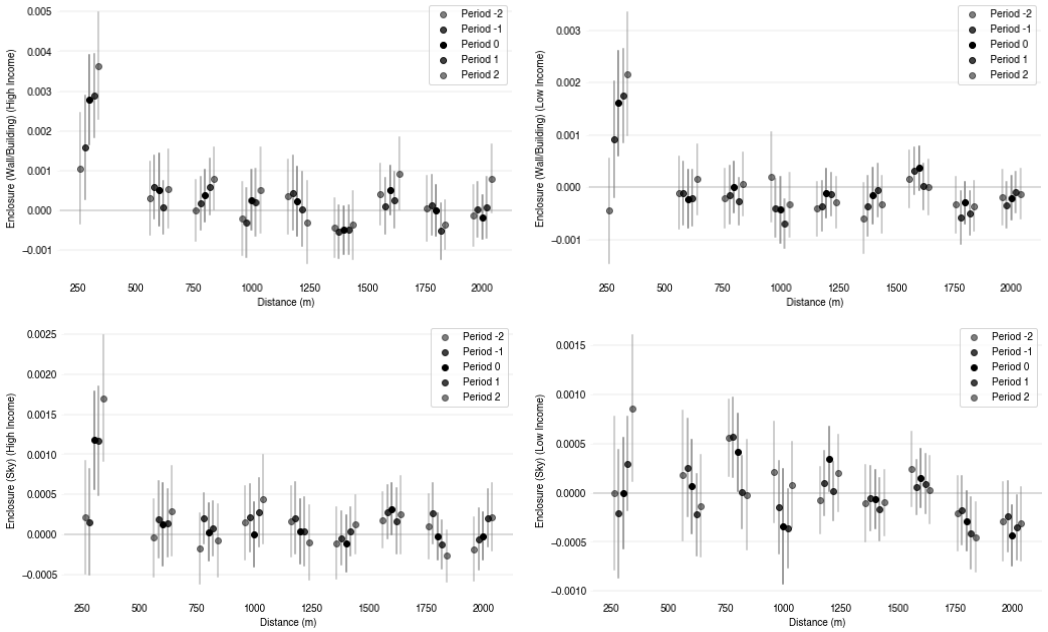

Figure C-7: Effects on enclosure in high- and low-income areas using the number of rooms as the treatment intensity measure

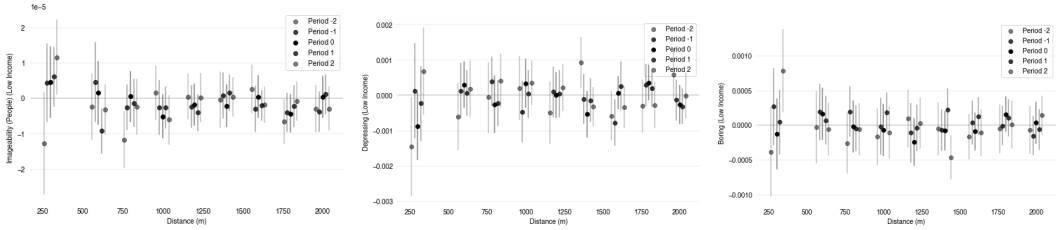

Figure C-8: Effects on imageability (people), depressing and boring in low-income areas using the number of units as the treatment intensity measure

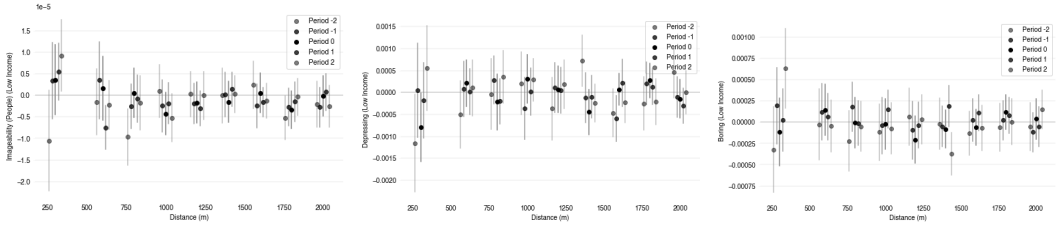

Figure C-9: Effects on imageability (people), depressing and boring in low-income areas using the number of rooms as the treatment intensity measure

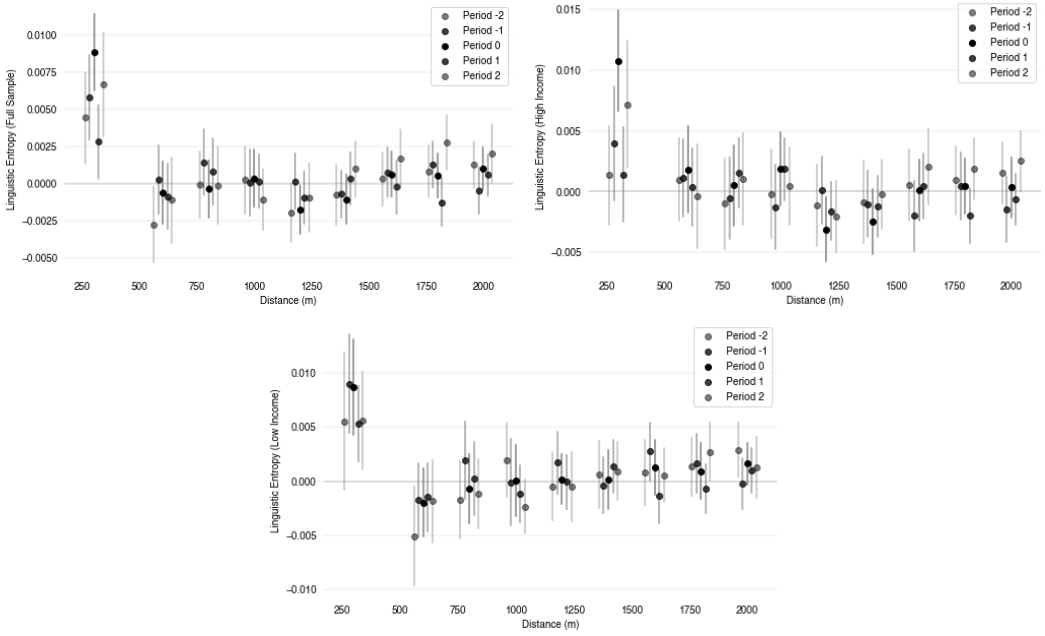

Figure C-10: Effects on linguistic entropy using the number of units as the treatment intensity measure

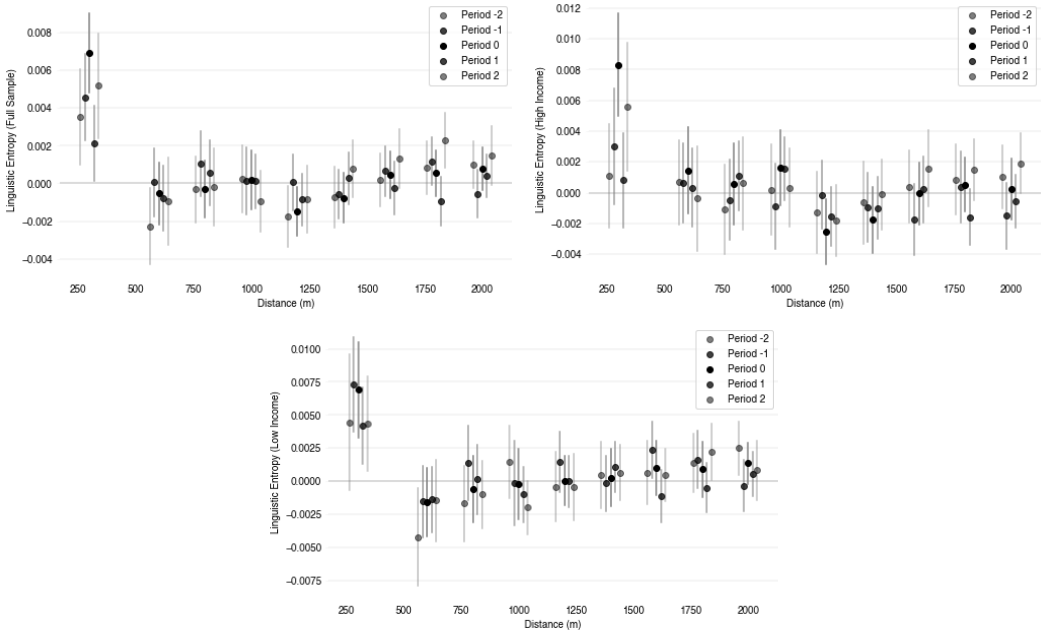

Figure C-11: Effects on linguistic entropy using the number of rooms as the treatment intensity measure

Bibliography

[1] Jan Amcoff. Searching for new ways to achieve mixed neighbourhoods. *Cities*, 121(103496), 2022.

[2] Yoshinobu Ashihara. *The aesthetic townscape*. MIT Press, 1983.

[3] Brian J Asquith, Evan Mast, and Davin Reed. Local effects of large new apartment buildings in low-income areas. *The Review of Economics and Statistics*, pages 1–46, 2021.

[4] Nationella Operativa Avdelningen. Utsatta områden - sociala risker, kollektiv förmåga och oönskade händelser, 2015.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[6] Monica Barni and Carla Bagna. The critical turn in LL: New methodologies and new items in LL. *Linguistic Landscape*, 1(1-2):6–18, 2015.

[7] Hector Blanco and Lorenzo Neri. Knocking it down and mixing it up: The impact of public housing regenerations. *MIT, Job Market Paper*, 2021.

[8] Bill Yang Cai, Xiaojiang Li, Ian Seiferling, and Carlo Ratti. Treepedia 2.0: Applying deep learning for large-scale quantification of urban tree cover. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 49–56. IEEE, 2018.

[9] Gerald A Carlino and Albert Saiz. Beautiful city: Leisure amenities and urban growth. *Journal of Regional Science*, 59(3):369–408, 2019.

[10] Long Chen, Yi Lu, Qiang Sheng, Yu Ye, Ruoyu Wang, and Ye Liu. Estimating pedestrian volume using street view images: A large-scale validation test. *Computers, Environment and Urban Systems*, 81:101481, 2020.

[11] Brett Christophers. A monstrous hybrid: The political economy of housing in early twenty-first century Sweden. *New Political Economy*, 18(6):885–911, 2013.

[12] City of Copenhagen. Good, better, best: The city of Copenhagen's bicycle strategy 2011-2025, 2011.

[13] City of Stockholm. Stockholm pedestrian plan, April 2016.

[14] City of Stockholm. Stockholm city plan, May 2018.

[15] Deborah A Cohen, Karen Mason, Ariane Bedimo, Richard Scribner, Victoria Basolo, and Thomas A Farley. Neighborhood physical conditions and health. *American Journal of Public Health*, 93(3):467–471, 2003.

[16] Gorden Cullen. The architectural press, 1961.

[17] Åke Daun. Swedish mentality. In *Swedish Mentality*. Penn State University Press, 2021.

[18] Rebecca Diamond and Tim McQuade. Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy*, 127(3):1063–1117, 2019.

[19] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 196–212, Cham, 2016. Springer International Publishing.

[20] Akilah Dulin-Keita, Herpreet Kaur Thind, Olivia Affuso, and Monica L Baskin. The associations of perceived neighborhood disorder and physical activity with obesity among African American adolescents. *BMC Public Health*, 13(1):1–10, 2013.

[21] Reid Ewing and Otto Clemente. *Measuring Urban Design: Metrics for Livable Places*. Island Press, 2013.

[22] Reid Ewing and Susan Handy. Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, 14(1):65–84, 2009.

[23] Jan Gehl. *Life between buildings*, volume 23. New York: Van Nostrand Reinhold, 1987.

[24] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137, 2018.

[25] Durk Gorter. Multilingual inequality in public spaces: towards an inclusive model of linguistic landscapes. *Multilingualism in the public space: Empowering and transforming communities*, 2021.

[26] Karin Grundström and Irene Molina. From folkhem to lifestyle housing in Sweden: segregation and urban form, 1930s–2010s. *International Journal of Housing Policy*, 16(3):316–336, 2016.

[27] Veronica Guerrieri, Daniel Hartley, and Erik Hurst. Endogenous gentrification and housing price dynamics. *Journal of Public Economics*, 100:45–60, 2013.

[28] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

[29] Thomas Hall and Sonja Vidén. The million homes programme: a review of the great swedish planning project. *Planning Perspectives*, 20(3):301–328, 2005.

[30] Heba Hassan, Ahmed El-Mahdy, and Mohamed E. Hussein. Arabic scene text recognition in the deep learning era: Analysis on a novel dataset. *IEEE Access*, 9:107046–107058, 2021.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[32] Karin Hedin, Eric Clark, Emma Lundholm, and Gunnar Malmberg. Neoliberalization of housing in Sweden: Gentrification, filtering, and social polarization. *Annals of the Association of American Geographers*, 102(2):443–463, 2012.

[33] Seong-Yun Hong. Linguistic landscapes on street-level images. *ISPRS International Journal of Geo-Information*, 9(1):57, 2020.

[34] Richard Hornbeck and Daniel Keniston. Creative destruction: Barriers to urban growth and the Great Boston Fire of 1872. *American Economic Review*, 107(6):1365–98, 2017.

[35] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[36] Francis M Hult. English on the streets of Sweden: An ecolinguistic view of two cities and a language policy. *Working Papers in Educational Linguistics (WPEL)*, 19(1):3, 2003.

[37] Allan B Jacobs. Looking at cities. *Places*, 1(4), 1984.

[38] Allan B Jacobs. *Great Streets*. Cambridge: MIT Press, 1993.

[39] Igor Knez, Åsa Ode Sang, Bengt Gunnarsson, and Marcus Hedblom. Wellbeing in urban greenery: the role of naturalness and place identity. *Frontiers in Psychology*, 9:491, 2018.

[40] Julia Koschinsky. Spatial heterogeneity in spillover effects of assisted and unassisted rental housing. *Journal of Urban Affairs*, 31(3):319–347, 2009.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[42] Mirte AG Kuipers, Mireille NM van Poppel, Wim van den Brink, Marleen Wingen, and Anton E Kunst. The association between neighborhood disorder, social cohesion and hazardous alcohol use: a national multilevel study. *Drug and Alcohol Dependence*, 126(1-2):27–34, 2012.

[43] Rodrigue Landry and Richard Y Bourhis. Linguistic landscape and ethnolinguistic vitality: An empirical study. *Journal of Language and Social Psychology*, 16(1):23–49, 1997.

[44] Johanna Leinonen et al. Researching in/visibility in the nordic context: Theoretical and empirical views. *Nordic Journal of Migration Research*, 4(4):161, 2014.

[45] Jaime Lerner. *Urban acupuncture*. Springer, 2014.

[46] Xiaodi Li. Do new housing units in your backyard raise your rents. *NYU Wagner and NYU Furman Center, Job Market Paper*, 57, 2019.

[47] Xiaojiang Li, Carlo Ratti, and Ian Seiferling. Quantifying the shade provision of street trees in urban landscape: A case study in Boston, USA, using Google Street View. *Landscape and Urban Planning*, 169:81–91, 2018.

[48] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685, 2015.

[49] Ruixian Ma, Wei Wang, Fan Zhang, Kyuha Shim, and Carlo Ratti. Typeface reveals spatial economical patterns. *Nature Scientific Reports*, 9(15946), 2019.

[50] Lars Marcus. Social housing and segregation in Sweden: from residential segregation to social integration in public space. *Progress in Planning*, 67(3):251–263, 2007.

[51] Ricardo Mora and Javier Ruiz-Castillo. Entropy-based segregation indices. *Sociological Methodology*, 41(1):159–194, 2011.

[52] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

[53] Kathryn M Neckerman, Marnie Purciel-Hill, James W Quinn, and Andrew Rundle. Urban design qualities for New York City. In *Measuring urban design*, pages 63–82. Springer, 2013.

[54] Laura Onofri, PALD Nunes, Jasone Cenoz, Durk Gorter, et al. Linguistic diversity and preferences: Econometric evidence from European cities. *Journal of Economics and Econometrics*, 56(1):39–60, 2013.

[55] Toril Opsahl. Invisible presence? polish in norwegian public spaces. *Multilingualism in Public Spaces: Empowering and Transforming Communities*, page 111, 2021.

[56] Vicente Ordonez and Tamara L Berg. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*, pages 494–510. Springer, 2014.

[57] Kate Pennington. Does building new housing cause displacement?: The supply and demand effects of construction in San Francisco. *UC Berkeley, Job Market Paper*, 2021.

[58] Douglas D Perkins, Courtney Larsen, and Barbara B Brown. Mapping urban revitalization: using GIS spatial analysis to evaluate a new housing policy. *Journal of Prevention & Intervention in the Community*, 37(1):48–65, 2009.

[59] Shanghai Planning and Land Resource Administration Bureau. *Shanghai Street Design Guidelines*. Tongji University Press, 2016.

[60] Esteban Rossi-Hansberg, Pierre-Daniel Sarte, and Raymond Owens III. Housing externalities. *Journal of Political Economy*, 118(3):485–535, 2010.

[61] Philip Salesses, Katja Schechtner, and César A Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS One*, 8(7):e68400, 2013.

[62] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

[63] Katharina Schone. The impact of inter-municipal land use plans on social segregation: first lessons from the french experience. *Local Government Studies*, pages 1–24, 2022.

[64] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2016.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[66] Barrett A Slade. Big-box stores and urban land prices: Friend or foe? *Real Estate Economics*, 46(1):7–58, 2018.

[67] Yan Song, Louis Merlin, and Daniel Rodriguez. Comparing measures of urban land use mix. *Computers, Environment and Urban Systems*, 42:1–13, 2013.

[68] Statistics Sweden. Open data for DeSO – demographic statistical areas, 2022.

[69] Bo-sin Tang and Kwan To Wong. Assessing externality: Successive event studies on market impacts of new housing development on an old residential neighbourhood. *Environment and Planning B: Urban Analytics and City Science*, 47(1):156–173, 2020.

[70] Jingxian Tang and Ying Long. Measuring visual quality of street space and its temporal variation: Methodology and its application in the hutong area in Beijing. *Landscape and Urban Planning*, 191:103436, 2019.

[71] Transport for London. Better streets delivered: Learning from completed schemes, September 2013.

[72] Yu Ye, Wei Zeng, Qiaomu Shen, Xiaohu Zhang, and Yi Lu. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. *Environment and Planning B: Urban Analytics and City Science*, 46(8):1439–1457, 2019.

[73] Fan Zhang, Zhuangyuan Fan, Yuhao Kang, Yujie Hu, and Carlo Ratti. "Perception bias": Deciphering a mismatch between urban crime and perception of safety. *Landscape and Urban Planning*, 207:104003, 2021.

[74] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160, 2018.

[75] Jianming Zhang, Chaoquan Lu, Xudong Li, Hye-Jin Kim, and Jin Wang. A full convolutional network based on densenet for remote sensing scene classification. *Mathematical Biosciences and Engineering*, 16(5):3345–3367, 2019.

[76] Ke Zhang, Yurong Guo, Xinsheng Wang, Jinsha Yuan, and Qiaolin Ding. Multiple feature reweight densenet for image classification. *IEEE Access*, 7:9872–9880, 2019.

[77] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.